

OPTIMIZATION OF NUMBER OF TRANSCEIVERS USED IN A SWITCH

FIELD OF THE INVENTION

The present invention is related to transferring fragments between port cards and fabrics. More specifically, the present invention is related to transferring fragments between port cards and fabrics with gigabit transceivers.

BACKGROUND OF THE INVENTION

The switch originally was designed to have high-speed HSTL point-to-point connections between port cards and fabrics running at 250 MHz. This was proven to be too fast and signal integrity could not be guaranteed due to the long trace length that exist. A more robust method had to be devised. The new method uses a network of gigabit transceivers between port cards and fabrics. The gigabit network is comprised of gigabit transmitters and receivers that communicate with each other at up to 1.3GHz. The interfaces to each transceiver are an 8-bit data bus running at 125 MHz. Since the frequency of operation was reduced by half, the bus widths had to be multiplied by two. This new method provides to be more robust and has better noise rejection since the gigabit transceivers use differential pairs to communicate with each other. Since these gigabit transceivers are replacing point-to-point buses and some buses are larger than 8 bits wide, multiple gigabit transceivers need to be used to make up these buses. The initial total number of gigabit transceivers required was as follows: 22 transmitters per Stripper, 40 receivers per Aggregator, 52 transmitters per Separator and 28 receivers per Unstriper. For a 480G switch, this would total to about (48 port cards)(4 Stripers)(22) + (13 fabrics)(9 Aggregators)(40) + (13 fabrics)(9 Separators)(52) + (48 port cards)(4 Unstripers)(28) = 4224 + 4680 + 6084 + 376 = 20364. This large number of gigabit transceivers

proved to consume too much power and take up too much boards space. At an average of 0.29 Watts of power dissipation per transceiver, the total power dissipation would have been 5,905 Watts $(20364)(.029)$.

5 The present invention using the gigabit transceivers solves the following problems:

1. Reduces the total number of gigabit transceivers required.
2. Reduces total power requirements of whole system.
3. Reduces board space requirements.
- 10 4. Handles un-synchronized updates to gigabit network.
5. Minimizes I/O requirements on backplane connectors.
6. Reduces PCB trace routing complexity.
7. Maximizes symmetry between ASICs which will simplify unique mux/demux structures required.

15 This optimization solves all the above problems by intelligently allocating gigabit transceivers to be re-used in modes where they can be re-used. This technique allows changes in the configuration of the switch (40G, 80G, 120G, etc.) To alter the gigabit transmitter and receiver assignment to be un-synchronized
20 as long as the switch configuration is changed up or down by one step in the configuration.

25 A switch which stripes data onto multiple fabrics and sends parity data to another fabric has been described in U.S. patent application serial number 09/333,450, incorporated by reference herein. See also U.S. patent application serial number 09/293,563 which describes a wide memory TDM switching system, incorporated by reference herein.

SUMMARY OF THE INVENTION

5 The present invention pertains to a switch. The switch comprises a plurality of port cards. The switch comprises a network connected to the port cards. Preferably, the network is a gigabit network. The gigabit network has transmitters and receivers that communicate with each other and have assignments between each other. Preferably, the transmitters and receivers are gigabit transmitters and gigabit receivers. The switch comprises a plurality of fabrics. Each fabric is connected to each port card through the gigabit network to send and receive fragments of packets to or from the port cards, the port cards, fabrics and network having a plurality of modes of operation. The switch comprises a control mechanism connected to the gigabit transmitters and gigabit receivers which changes the assignments according to the mode. The control mechanism changes the mode and reuses the gigabit transmitters and gigabit receivers where they can be reused.

10
15
20 The present invention pertains to a method for switching fragments of packets. The method comprises the steps of assigning assignments between transmitters and receivers of a network. Preferably, the network is a gigabit network. Preferably, the transmitters and receivers are gigabit transmitters and gigabit receivers. Next there is the step of changing a mode of the fabrics, port cards and networks. Then there is the step of changing the assignments of the transmitters and receivers according to the mode and reusing the gigabit transmitters and gigabit receivers where they can be reused. Then there is the step of transferring the fragments of packets between fabrics and port cards with the gigabit transmitters and gigabit receivers of the gigabit network.

BRIEF DESCRIPTION OF THE DRAWINGS

In the accompanying drawings, the preferred embodiment of the invention and preferred methods of practicing the invention are illustrated in which:

5 Figure 1 is a schematic representation of packet striping in the switch of the present invention.

 Figure 2 is a schematic representation of an OC 48 port card.

10 Figure 3 is a schematic representation of a concatenated network blade.

 Figure 4 is a schematic representation regarding the connectivity of the fabric ASICs.

 Figure 5 is a schematic representation of a 32-bit cell transfer.

15 Figure 6 is a schematic representation regarding back-pressure.

 Figure 7 is a schematic representation of a 32-bit packet transferred using external connection number bus.

20 Figure 8 is a schematic representation of a 64-bit cell transferred.

 Figure 9 is a schematic representation of a 64-bit packet transfer.

Figure 10 is a schematic representation of ATM cell flow in the switch.

Figure 11 is a schematic representation of sync pulse distribution.

5 Figure 12 is a schematic representation regarding the write cycle.

Figure 13 is a schematic representation of the read cycle.

10 Figure 14 is a schematic representation of the striper ASIC architecture.

Figure 15 is a schematic presentation of the aggregator ASIC architecture.

Figure 16 is a schematic representation of a memory controller ASIC architecture.

15 Figure 17 is a schematic representation of the wide cache line shared memory architecture.

Figure 18 is a schematic representation of a separator ASIC architecture.

20 Figure 19 is a schematic representation of an unstriper ASIC architecture.

Figure 20 is a schematic representation regarding the relationship between transmit and receive sequence counters for the separator and unstriper, respectively.

Figure 21 is a schematic representation of a receive synchronizer.

Figure 22 is a schematic representation of a switch of the present invention.

5

DETAILED DESCRIPTION

Referring now to the drawings wherein like reference numerals refer to similar or identical parts throughout the several views, and more specifically to figure 22 thereof, there is shown a switch 10. The switch 10 comprises at least one port card 12. The switch 10 comprises a network 14 connected to the port card 12. Preferably, the network is a gigabit network. The gigabit network 14 has transmitters 16 and receivers 18 that communicate with each other and have assignments between each other. Preferably, the transmitters and receivers are gigabit transmitters and gigabit receivers. The switch 10 comprises at least one fabric 20. Each fabric 20 is connected to each port card 12 through the gigabit network 14 to send and receive fragments of packets to or from the port cards 12, the port cards 12, fabrics 20 and network 14 having a plurality of modes of operation. The switch 10 comprises a control mechanism 22 connected to the gigabit transmitters 16 and gigabit receivers 18 which changes the assignments according to the mode. The control mechanism 22 changes the mode and reuses the gigabit transmitters 16 and gigabit receivers 18 where they can be reused.

25

Preferably, the network 14 includes a mux structure 24 that makes the assignments between transmitters 16 and receivers 18. The modes preferably are 40G, 80G, 120G, 240G slow, 240G fast or 480G. Preferably, the network 14 can support two modes simultaneously as combined modes. The combined modes preferably

are 40/80, 80/120, 120/160, 160/240 fast, 240 fast/240 slow, or 240 slow/480.

Preferably, each port card 12 includes a striper 26 and an unstriper 28. Each fabric 20 preferably includes a separator 30 and an aggregator 32. Preferably, the assignments between transmitters 16 and receivers 18 communicate with each other at up to 1.3 GHz. Each transmitter 16 preferably takes in 8 bits of data and 2 bits of control and serially transmits the bits of data and control to the associated receiver. Preferably, each receiver recovers clock and data it receives by using an 8B/10B decoding protocol and provides 8 bits of data and 3 bits of control.

The present invention pertains to a method for switching fragments of packets. The method comprises the steps of assigning assignments between transmitters 16 and receivers 18 of a network 14. Preferably, the network is a gigabit network. Preferably, the transmitters and receivers are gigabit transmitters and gigabit receivers. Next there is the step of changing a mode of the fabrics 20, port cards 12 and networks 14. Then there is the step of changing the assignments of the transmitters 16 and receivers 18 according to the mode and reusing the gigabit transmitters 16 and gigabit receivers 18 where they can be reused. Then there is the step of transferring the fragments of packets between fabrics 20 and port cards 12 with the gigabit transmitters 16 and gigabit receivers 18 of the gigabit network 14.

Preferably, the changing of the assignments step includes the step of rearranging the mux structure 24 of the network 14 that makes the assignments between the receivers 18 and transmitters 16. The changing the mode step preferably includes the step of changing the mode between 40G and 80G, or 80G and 120G, or 120G and 160G or 160G and 240G slow or 240G slow and 240G fast. Preferably, the

assigning step includes the step of assigning assignments between the receivers 18 and transmitters 16 to support two modes simultaneously as combined modes. The transferring step preferably includes the step of transferring with the transmitter 8 bits of data and 2 bits of control serially through a mux structure 24 to the receiver assigned to the transmitter.

In the operation of the invention, the switch uses a network 14 of gigabit transceivers that make the connections from port cards 12 to fabrics 20 across the backplane. The gigabit network 14 is comprised of gigabit transmitters 16 and receivers 18 that communicate with each other at up to 1.3 GHz by using differential pairs. Each gigabit transmitter 16 takes in 8 bits of data and 2 bits of control and serially transmits this data to a gigabit receiver that recovers clock and data and by using an 8B/10B decoding protocol, provides 8 bits of data and 3 bits of control. Between the transmitters 16 and receivers 18 exists a mux structure 24 that makes the assignment of transmitters 16 to receivers 18. The mux structure 24 can be configured by software control to re-assign transceivers based on the mode of the switch 10. The mux structure 24 utilizes built-in 2-to-1 muxes in each transceiver. The mux structure 24 could not be built entirely with 2-to-1 muxes since in some cases 3-to-1 muxes were required. Therefore, some additional external 2-to-1 muxes are required. The seven possible configurations are as follows 40G, 80G, 120G, 160G, 240G slow, 240G fast, 480G. Since updates to the gigabit network 14 cannot be synchronized across the whole switch 10, it was decided to combine the switch 10 configurations in pairs so that at any particular time, the gigabit network 14 would support two configurations simultaneously. This can be done because pairs of configuration modes have transceivers that are common to both and

can be shared. However, the switch 10 configuration is restricted to move up or down in capacity by one mode at a time. The combined modes that are supported as follows: 40/80, 80/120, 120/160, 160/240 fast, 240 fast/240 slow, 240 slow/480. 240 fast and slow
5 modes are described later.

10 The minimum requirement on the number of gigabit transceivers can be obtained by looking at the minimum number of independent channels that need to be supported by each ASIC. For the Striper 26 and Unstriper 28, 13 independent sources and destinations exist respectively, that need their own clock source
15 in 480G mode. The Striper 26 needs to source 2 routeward bits, 2 data bits and a back-pressure bit for a total of 5 bits per transmitter used. The Separator 30 needs to source 4 routeward bits and 4 data bits for a combined total of 8 bits per
20 transmitter. Each transceiver can handle 8 bits of data so one transceiver per channel in 480 mode is sufficient. This limits the reduction of gigabit transceivers to at least 13 per striper and unstriper since each gigabit transceiver provides only one clock source. A similar requirement holds true for the Aggregator 32 and
25 Separator 30. The Aggregator 32 and Separator 30 require 24 independent channels in 480 mode and therefore require a minimum of 24 gigabit transceivers each. After some analysis, it was shown that the original speedup of 2 on the egress of the switch was compromised in some modes. The speedup for 240 mode was reduced to
30 1 and was not acceptable. To have a speedup of 2 in 240 mode, one additional transceiver was required for the Unstriper 28. To make things symmetrical, an additional transmitter 16 was added to the Striper 26. 240 mode was split into two modes. 240 fast mode has a speedup of 2 and 240 slow mode has a speedup of 1. 240 slow mode needs to also be used since no common mapping of transceivers could be found from 240 fast mode to 480 mode. This is true since both modes use all 14 transceivers but have completely different

mappings. This can be seen in the Appendix "BFS Backplane Optimization Worksheet". This shows how the gigabit transceiver assignments were derived. The original map was done for the Unstriper 28 and from this, it can be seen that in 240 fast mode, all 14 transceivers were required. All other modes were assigned based on the 240 fast mode and unused transceivers were re-used in modes that required them. The Striper 26 map is a subset of the Unstriper 28 map and both Aggregator 32 maps and the Separator 30 5-8 map are derived from the Separator 30 1-4 map. In the Appendix, the "BFS ASIC Backplane Connection Map" shows the final gigabit transceiver assignments. This is a summary of the mappings derived earlier. The document in the Appendix "Gigabit Signals", shows how the number of data bits per bus for each mode affect the number of gigabit transceivers used.

After this optimization, the total gigabit transceiver requirement is as follows: 14 transmitters per Striper, 24 receivers 18 per Aggregator, 24 transmitters per Separator and 14 receivers per Unstriper 28. For a 480G switch, this would total to about (48 port cards)(4 Stripers)(14) + (13 fabrics)(9 Aggregators)(24) + (13 fabrics)(9 Separators)(24) + (48 port cards)(4 Unstripers)(14) = 2688 + 2808 + 2808 + 2688 = 10992. This is a 46% reduction in the number of transceivers required. At an average of 0.29 Watts of power dissipation per transceiver, the total power dissipation will be 3,187 Watts (10992)(0.29).

The switch uses RAID techniques to increase overall switch bandwidth while minimizing individual fabric bandwidth. In the switch architecture, all data is distributed evenly across all fabrics so the switch adds bandwidth by adding fabrics and the fabric need not increase its bandwidth capacity as the switch increases bandwidth capacity.

Each fabric provides 40G of switching bandwidth and the system supports 1, 2, 3, 4, 6, or 12 fabrics, exclusive of the redundant/spare fabric. In other words, the switch can be a 40G, 80G, 120G, 160G, 240G, or 480G switch depending on how many fabrics
5 are installed.

A portcard provides 10G of port bandwidth. For every 4 portcards, there needs to be 1 fabric. The switch architecture does not support arbitrary installations of portcards and fabrics.

0 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390 395 400 405 410 415 420 425 430 435 440 445 450 455 460 465 470 475 480 485 490 495 500 505 510 515 520 525 530 535 540 545 550 555 560 565 570 575 580 585 590 595 600 605 610 615 620 625 630 635 640 645 650 655 660 665 670 675 680 685 690 695 700 705 710 715 720 725 730 735 740 745 750 755 760 765 770 775 780 785 790 795 800 805 810 815 820 825 830 835 840 845 850 855 860 865 870 875 880 885 890 895 900 905 910 915 920 925 930 935 940 945 950 955 960 965 970 975 980 985 990 995 1000 1005 1010 1015 1020 1025 1030 1035 1040 1045 1050 1055 1060 1065 1070 1075 1080 1085 1090 1095 1100 1105 1110 1115 1120 1125 1130 1135 1140 1145 1150 1155 1160 1165 1170 1175 1180 1185 1190 1195 1200 1205 1210 1215 1220 1225 1230 1235 1240 1245 1250 1255 1260 1265 1270 1275 1280 1285 1290 1295 1300 1305 1310 1315 1320 1325 1330 1335 1340 1345 1350 1355 1360 1365 1370 1375 1380 1385 1390 1395 1400 1405 1410 1415 1420 1425 1430 1435 1440 1445 1450 1455 1460 1465 1470 1475 1480 1485 1490 1495 1500 1505 1510 1515 1520 1525 1530 1535 1540 1545 1550 1555 1560 1565 1570 1575 1580 1585 1590 1595 1600 1605 1610 1615 1620 1625 1630 1635 1640 1645 1650 1655 1660 1665 1670 1675 1680 1685 1690 1695 1700 1705 1710 1715 1720 1725 1730 1735 1740 1745 1750 1755 1760 1765 1770 1775 1780 1785 1790 1795 1800 1805 1810 1815 1820 1825 1830 1835 1840 1845 1850 1855 1860 1865 1870 1875 1880 1885 1890 1895 1900 1905 1910 1915 1920 1925 1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020 2025 2030 2035 2040 2045 2050 2055 2060 2065 2070 2075 2080 2085 2090 2095 2100 2105 2110 2115 2120 2125 2130 2135 2140 2145 2150 2155 2160 2165 2170 2175 2180 2185 2190 2195 2200 2205 2210 2215 2220 2225 2230 2235 2240 2245 2250 2255 2260 2265 2270 2275 2280 2285 2290 2295 2300 2305 2310 2315 2320 2325 2330 2335 2340 2345 2350 2355 2360 2365 2370 2375 2380 2385 2390 2395 2400 2405 2410 2415 2420 2425 2430 2435 2440 2445 2450 2455 2460 2465 2470 2475 2480 2485 2490 2495 2500 2505 2510 2515 2520 2525 2530 2535 2540 2545 2550 2555 2560 2565 2570 2575 2580 2585 2590 2595 2600 2605 2610 2615 2620 2625 2630 2635 2640 2645 2650 2655 2660 2665 2670 2675 2680 2685 2690 2695 2700 2705 2710 2715 2720 2725 2730 2735 2740 2745 2750 2755 2760 2765 2770 2775 2780 2785 2790 2795 2800 2805 2810 2815 2820 2825 2830 2835 2840 2845 2850 2855 2860 2865 2870 2875 2880 2885 2890 2895 2900 2905 2910 2915 2920 2925 2930 2935 2940 2945 2950 2955 2960 2965 2970 2975 2980 2985 2990 2995 3000 3005 3010 3015 3020 3025 3030 3035 3040 3045 3050 3055 3060 3065 3070 3075 3080 3085 3090 3095 3100 3105 3110 3115 3120 3125 3130 3135 3140 3145 3150 3155 3160 3165 3170 3175 3180 3185 3190 3195 3200 3205 3210 3215 3220 3225 3230 3235 3240 3245 3250 3255 3260 3265 3270 3275 3280 3285 3290 3295 3300 3305 3310 3315 3320 3325 3330 3335 3340 3345 3350 3355 3360 3365 3370 3375 3380 3385 3390 3395 3400 3405 3410 3415 3420 3425 3430 3435 3440 3445 3450 3455 3460 3465 3470 3475 3480 3485 3490 3495 3500 3505 3510 3515 3520 3525 3530 3535 3540 3545 3550 3555 3560 3565 3570 3575 3580 3585 3590 3595 3600 3605 3610 3615 3620 3625 3630 3635 3640 3645 3650 3655 3660 3665 3670 3675 3680 3685 3690 3695 3700 3705 3710 3715 3720 3725 3730 3735 3740 3745 3750 3755 3760 3765 3770 3775 3780 3785 3790 3795 3800 3805 3810 3815 3820 3825 3830 3835 3840 3845 3850 3855 3860 3865 3870 3875 3880 3885 3890 3895 3900 3905 3910 3915 3920 3925 3930 3935 3940 3945 3950 3955 3960 3965 3970 3975 3980 3985 3990 3995 4000 4005 4010 4015 4020 4025 4030 4035 4040 4045 4050 4055 4060 4065 4070 4075 4080 4085 4090 4095 4100 4105 4110 4115 4120 4125 4130 4135 4140 4145 4150 4155 4160 4165 4170 4175 4180 4185 4190 4195 4200 4205 4210 4215 4220 4225 4230 4235 4240 4245 4250 4255 4260 4265 4270 4275 4280 4285 4290 4295 4300 4305 4310 4315 4320 4325 4330 4335 4340 4345 4350 4355 4360 4365 4370 4375 4380 4385 4390 4395 4400 4405 4410 4415 4420 4425 4430 4435 4440 4445 4450 4455 4460 4465 4470 4475 4480 4485 4490 4495 4500 4505 4510 4515 4520 4525 4530 4535 4540 4545 4550 4555 4560 4565 4570 4575 4580 4585 4590 4595 4600 4605 4610 4615 4620 4625 4630 4635 4640 4645 4650 4655 4660 4665 4670 4675 4680 4685 4690 4695 4700 4705 4710 4715 4720 4725 4730 4735 4740 4745 4750 4755 4760 4765 4770 4775 4780 4785 4790 4795 4800 4805 4810 4815 4820 4825 4830 4835 4840 4845 4850 4855 4860 4865 4870 4875 4880 4885 4890 4895 4900 4905 4910 4915 4920 4925 4930 4935 4940 4945 4950 4955 4960 4965 4970 4975 4980 4985 4990 4995 5000 5005 5010 5015 5020 5025 5030 5035 5040 5045 5050 5055 5060 5065 5070 5075 5080 5085 5090 5095 5100 5105 5110 5115 5120 5125 5130 5135 5140 5145 5150 5155 5160 5165 5170 5175 5180 5185 5190 5195 5200 5205 5210 5215 5220 5225 5230 5235 5240 5245 5250 5255 5260 5265 5270 5275 5280 5285 5290 5295 5300 5305 5310 5315 5320 5325 5330 5335 5340 5345 5350 5355 5360 5365 5370 5375 5380 5385 5390 5395 5400 5405 5410 5415 5420 5425 5430 5435 5440 5445 5450 5455 5460 5465 5470 5475 5480 5485 5490 5495 5500 5505 5510 5515 5520 5525 5530 5535 5540 5545 5550 5555 5560 5565 5570 5575 5580 5585 5590 5595 5600 5605 5610 5615 5620 5625 5630 5635 5640 5645 5650 5655 5660 5665 5670 5675 5680 5685 5690 5695 5700 5705 5710 5715 5720 5725 5730 5735 5740 5745 5750 5755 5760 5765 5770 5775 5780 5785 5790 5795 5800 5805 5810 5815 5820 5825 5830 5835 5840 5845 5850 5855 5860 5865 5870 5875 5880 5885 5890 5895 5900 5905 5910 5915 5920 5925 5930 5935 5940 5945 5950 5955 5960 5965 5970 5975 5980 5985 5990 5995 6000 6005 6010 6015 6020 6025 6030 6035 6040 6045 6050 6055 6060 6065 6070 6075 6080 6085 6090 6095 6100 6105 6110 6115 6120 6125 6130 6135 6140 6145 6150 6155 6160 6165 6170 6175 6180 6185 6190 6195 6200 6205 6210 6215 6220 6225 6230 6235 6240 6245 6250 6255 6260 6265 6270 6275 6280 6285 6290 6295 6300 6305 6310 6315 6320 6325 6330 6335 6340 6345 6350 6355 6360 6365 6370 6375 6380 6385 6390 6395 6400 6405 6410 6415 6420 6425 6430 6435 6440 6445 6450 6455 6460 6465 6470 6475 6480 6485 6490 6495 6500 6505 6510 6515 6520 6525 6530 6535 6540 6545 6550 6555 6560 6565 6570 6575 6580 6585 6590 6595 6600 6605 6610 6615 6620 6625 6630 6635 6640 6645 6650 6655 6660 6665 6670 6675 6680 6685 6690 6695 6700 6705 6710 6715 6720 6725 6730 6735 6740 6745 6750 6755 6760 6765 6770 6775 6780 6785 6790 6795 6800 6805 6810 6815 6820 6825 6830 6835 6840 6845 6850 6855 6860 6865 6870 6875 6880 6885 6890 6895 6900 6905 6910 6915 6920 6925 6930 6935 6940 6945 6950 6955 6960 6965 6970 6975 6980 6985 6990 6995 7000 7005 7010 7015 7020 7025 7030 7035 7040 7045 7050 7055 7060 7065 7070 7075 7080 7085 7090 7095 7100 7105 7110 7115 7120 7125 7130 7135 7140 7145 7150 7155 7160 7165 7170 7175 7180 7185 7190 7195 7200 7205 7210 7215 7220 7225 7230 7235 7240 7245 7250 7255 7260 7265 7270 7275 7280 7285 7290 7295 7300 7305 7310 7315 7320 7325 7330 7335 7340 7345 7350 7355 7360 7365 7370 7375 7380 7385 7390 7395 7400 7405 7410 7415 7420 7425 7430 7435 7440 7445 7450 7455 7460 7465 7470 7475 7480 7485 7490 7495 7500 7505 7510 7515 7520 7525 7530 7535 7540 7545 7550 7555 7560 7565 7570 7575 7580 7585 7590 7595 7600 7605 7610 7615 7620 7625 7630 7635 7640 7645 7650 7655 7660 7665 7670 7675 7680 7685 7690 7695 7700 7705 7710 7715 7720 7725 7730 7735 7740 7745 7750 7755 7760 7765 7770 7775 7780 7785 7790 7795 7800 7805 7810 7815 7820 7825 7830 7835 7840 7845 7850 7855 7860 7865 7870 7875 7880 7885 7890 7895 7900 7905 7910 7915 7920 7925 7930 7935 7940 7945 7950 7955 7960 7965 7970 7975 7980 7985 7990 7995 8000 8005 8010 8015 8020 8025 8030 8035 8040 8045 8050 8055 8060 8065 8070 8075 8080 8085 8090 8095 8100 8105 8110 8115 8120 8125 8130 8135 8140 8145 8150 8155 8160 8165 8170 8175 8180 8185 8190 8195 8200 8205 8210 8215 8220 8225 8230 8235 8240 8245 8250 8255 8260 8265 8270 8275 8280 8285 8290 8295 8300 8305 8310 8315 8320 8325 8330 8335 8340 8345 8350 8355 8360 8365 8370 8375 8380 8385 8390 8395 8400 8405 8410 8415 8420 8425 8430 8435 8440 8445 8450 8455 8460 8465 8470 8475 8480 8485 8490 8495 8500 8505 8510 8515 8520 8525 8530 8535 8540 8545 8550 8555 8560 8565 8570 8575 8580 8585 8590 8595 8600 8605 8610 8615 8620 8625 8630 8635 8640 8645 8650 8655 8660 8665 8670 8675 8680 8685 8690 8695 8700 8705 8710 8715 8720 8725 8730 8735 8740 8745 8750 8755 8760 8765 8770 8775 8780 8785 8790 8795 8800 8805 8810 8815 8820 8825 8830 8835 8840 8845 8850 8855 8860 8865 8870 8875 8880 8885 8890 8895 8900 8905 8910 8915 8920 8925 8930 8935 8940 8945 8950 8955 8960 8965 8970 8975 8980 8985 8990 8995 9000 9005 9010 9015 9020 9025 9030 9035 9040 9045 9050 9055 9060 9065 9070 9075 9080 9085 9090 9095 9100 9105 9110 9115 9120 9125 9130 9135 9140 9145 9150 9155 9160 9165 9170 9175 9180 9185 9190 9195 9200 9205 9210 9215 9220 9225 9230 9235 9240 9245 9250 9255 9260 9265 9270 9275 9280 9285 9290 9295 9300 9305 9310 9315 9320 9325 9330 9335 9340 9345 9350 9355 9360 9365 9370 9375 9380 9385 9390 9395 9400 9405 9410 9415 9420 9425 9430 9435 9440 9445 9450 9455 9460 9465 9470 9475 9480 9485 9490 9495 9500 9505 9510 9515 9520 9525 9530 9535 9540 9545 9550 9555 9560 9565 9570 9575 9580 9585 9590 9595 9600 9605 9610 9615 9620 9625 9630 9635 9640 9645 9650 9655 9660 9665 9670 9675 9680 9685 9690 9695 9700 9705 9710 9715 9720 9725 9730 9735 9740 9745 9750 9755 9760 9765 9770 9775 9780 9785 9790 9795 9800 9805 9810 9815 9820 9825 9830 9835 9840 9845 9850 9855 9860 9865 9870 9875 9880 9885 9890 9895 9900 9905 9910 9915 9920 9925 9930 9935 9940 9945 9950 9955 9960 9965 9970 9975 9980 9985 9990 9995 10000 10005 10010 10015 10020 10025 10030 10035 10040 10045 10050 10055 10060 10065 10070 10075 10080 10085 10090 10095 10100 10105 10110 10115 10120 10125 10130 10135 10140 10145 10150 10155 10160 10165 10170 10175 10180 10185 10190 10195 10200 10205 10210 10215 10220 10225 10230 10235 10240 10245 10250 10255 10260 10265 10270 10275 10280 10285 10290 10295 10300 10305 10310 10315 10320 10325 10330 10335 10340 10345 10350 10355 10360 10365 10370 10375 10380 10385 10390 10395 10400 10405 10410 10415 10420 10425 10430 10435 10440 10445 10450 10455 10460 10465 10470 10475 10480 10485 10490 10495 10500 10505 10510 10515 10520 10525 10530 10535 10540 10545 10550 10555 10560 10565 10570 10575 10580 10585 10590 10595 10600 10605 10610 10615 10620 10625 10630 10635 10640 10645 10650 10655 10660 10665 10670 10675 10680 10685 10690 10695 10700 10705 10710 10715 10720 10725 10730 10735 10740 10745 10750 10755 10760 10765 10770 10775 10780 10785 10790 10795 10800 10805 10810 10815 10820 10825 10830 10835 10840 10845 10850 10855 10860 10865 10870 10875 10880 10885 10890 10895 10900 10905 10910 10915 10920 10925 10930 10935 10940 10945 10950 10955 10960 10965 10970 10975 10980 10985 10990 10995 11000 11005 11010 11015 11020 11025 11030 11035 11040 11045 11050 11055 11060 11065 11070 11075 11080 11085 11090 11095 11100 11105 11110 11115 11120 11125 11130 11135 11140 11145 11150 11155 11160 11165 11170 11175 11180 11185 11190 11195 11200 11205 11210 11215 11220 11225 11230 11235 11240 11245 11250 11255 11260 11265 11270 11275 11280 11285 11290 11295 11300 11305 11310 11315 11320 11325 11330 11335 11340 11345 11350 11355 11360 11365 11370 11375 11380 11385 11390 11395 11400 11405 11410 11415 11420 11425 11430 11435 11440 11445 11450 11455 11460 11465 11470 11475 11480 11485 11490 11495 11500 11505 11510 11515 11520 11525 11530 11535 11540 11545 11550 11555 11560 11565 11570 11575 11580 11585 11590 11595 11600 11605 11610 11615 11620 11625 11630 11635 11640 11645 11650 11655 11660 11665 11670 11675 11680 11685 11690 11695 11700 11705 11710 11715 11720 11725 11730 11735 11740 11745 11750 11755 11760 11765 11770 11775 11780 11785 11790 11795 11800 11805 11810 11815 11820 11825 11830 11835 11840 11845 11850 11855 11860 11865 11870 11875 11880 11885 11890 11895 11900 11905 11910 11915 11920 11925 11930 11935 11940 11945 119

3. Aggregator - The Aggregator takes the data streams and routewords from the Stripers and multiplexes them into a single input stream to the Memory Controller.
4. Memory Controller - The Memory controller implements the queueing and dequeueing mechanisms of the switch. This includes the proprietary wide memory interface to achieve the simultaneous en-/de-queueing of multiple cells of data per clock cycle. The dequeueing side of the Memory Controller runs at 80Gbps compared to 40Gbps in order to make the bulk of the queueing and shaping of connections occur on the portcards.
5. Separator - The Separator implements the inverse operation of the Aggregator. The data stream from the Memory Controller is demultiplexed into multiple streams of data and forwarded to the appropriate Unstriper ASIC. Included in the interface to the Unstriper is a queue and flow control handshaking.
6. Trident - Trident is, strictly speaking, not one of the ASICs. It is actually one-half of the Poseidon chipset. Trident will be used to implement the ATM portcards within the switch.
7. Vortex - Vortex is the partner to Trident in the Poseidon chipset. Vortex is the ingress ASIC and Trident the egress device. Together, the two chips

implement a 2.5Gbps ingress, 5Gbps egress system capable of supporting up to OC-48c ports.

8. Reassembler - The Reassembler ASIC is the frame blade equivalent to Trident. It will be capable of taking cell streams from the Unstriper and converting them into frames.

There are 3 different views one can take of the connections between the fabric: physical, logical, and "active." Physically, the connections between the portcards and the fabrics are all gigabit speed differential pair serial links. This is strictly an implementation issue to reduce the number of signals going over the backplane. The "active" perspective looks at a single switch configuration, or it may be thought of as a snapshot of how data is being processed at a given moment. The interface between the fabric ASIC on the portcards and the fabrics is effectively 12 bits wide. Those 12 bits are evenly distributed ("striped") across 1, 2, 3, 4, 6, or 12 fabrics based on how the fabric ASICs are configured. The "active" perspective refers to the number of bits being processed by each fabric in the current configuration which is exactly 12 divided by the number of fabrics.

The logical perspective can be viewed as the union or max function of all the possible active configurations. Fabric slot #1 can, depending on configuration, be processing 12, 6, 4, 3, 2, or 1 bits of the data from a single Striper and is therefore drawn with a 12 bit bus. In contrast, fabric slot #3 can only be used to process 4, 3, 2, or 1 bits from a single Striper and is therefore drawn with a 4 bit bus.

Unlike previous switches, the switch really doesn't have a concept of a software controllable fabric redundancy mode. The

fabric ASICs implement N+1 redundancy without any intervention as long as the spare fabric is installed.

As far as what does it provide; N+1 redundancy means that the hardware will automatically detect and correct a single failure without the loss of any data.

The way the redundancy works is fairly simple, but to make it even simpler to understand a specific case of a 120G switch is used which has 3 fabrics (A, B, and C) plus a spare (S). The Striper takes the 12 bit bus and first generates a checkword which gets appended to the data unit (cell or frame). The data unit and checkword are then split into a 4-bit-per-clock-cycle data stripe for each of the A, B, and C fabrics ($A_3A_2A_1A_0$, $B_3B_2B_1B_0$, and $C_3C_2C_1C_0$). These stripes are then used to produce the stripe for the spare fabric $S_3S_2S_1S_0$ where $S_n = A_n \text{ XOR } B_n \text{ XOR } C_n$ and these 4 stripes are sent to their corresponding fabrics. On the other side of the fabrics, the Unstriper receives 4 4-bit stripes from A, B, C, and S. All possible combinations of 3 fabrics (ABC, ABS, ASC, and SBC) are then used to reconstruct a "tentative" 12-bit data stream. A checkword is then calculated for each of the 4 tentative streams and the calculated checkword compared to the checkword at the end of the data unit. If no error occurred in transit, then all 4 streams will have checkword matches and the ABC stream will be forwarded to the Unstriper output. If a (single) error occurred, only one checkword match will exist and the stream with the match will be forwarded off chip and the Unstriper will identify the faulty fabric stripe.

For different switch configurations, i.e. 1, 2, 4, 6, or 12 fabrics, the algorithm is the same but the stripe width changes.

If 2 fabrics fail, all data running through the switch will almost certainly be corrupted.

There are basically two options, both requiring that the defective fabrics be known through some means. Unfortunately, in a double failure system, the hardware that detects and identifies a failed fabric will only be able to identify the fabric that failed first (if there was one). Identifying both the failed fabrics may only be possible through a trial-and-error approach unless the switch software and/or switch diagnostics can develop tests to identify the second failure.

The recommended approach would be to shut down the switch and install as many good fabrics as possible beginning with slot 1. This allows the maximum bandwidth and redundancy be available given the functional hardware available.

The other option is to have the switch software reconfigure the switch to use fewer fabrics. This is an inferior solution for two reasons:

1. It can never provide more bandwidth than the recommended approach.
2. It requires substantial thought and understanding of the switch by the user in order to determine what is the maximum operational configuration.

Basically, the user must start at fabric slot 1 and count the number of operational fabrics. If the spare fabric is operational, then it may be used to "cover" for the first non-operational fabrics.

Example #1: A redundant 240G switch (6+1 fabrics) has suffered fabric failures in slots 3 and 4. Starting with slot 1 there are 2 operational fabrics and the spare is available to cover for slot 3. This switch can be reconfigured to a 120G non-redundant switch or an 80G redundant switch. Note than by swapping fabric 5 and 6 into slots 3 and 4, this switch could be a 160G redundant switch.

Example #2: A redundant 480G switch suffers fabric failures in slots 1 and the spare. Start swapping fabrics. Slot 1 is dead and the spare is not available to cover for it. This is the worst case scenario.

Example #3: A redundant 480G switch suffers fabric failures in slots 2 and 10. There is one functional fabric counting from slot 1 or 9 if the spare is used to cover for slot 2. This switch can be configured either as 40G redundant or 240G non-redundant. Note that fabrics 7,8, and 9 do not help since the only legal configuration after 6 fabrics is all 12.

The fabric slots are numbered and must be populated in ascending order. Also, the spare fabric is a specific slot so populating fabric slots 1, 2, 3, and 4 is different than populating fabric slots 1, 2, 3, and the spare. The former is a 160G switch without redundancy and the latter is 120G with redundancy.

Firstly, the ASICs are constructed and the backplane connected such that the use of a certain portcard slots requires there to be at least a certain minimum number of fabrics installed, not including the spare. This relationship is shown in Table 0.

In addition, the APS redundancy within the switch is limited to specifically paired portcards. Portcards 1 and 2 are paired, 3 and 4 are paired, and so on through portcards 47 and 48. This means that if APS redundancy is required, the paired slots must be populated together.

To give a simple example, take a configuration with 2 portcards and only 1 fabric. If the user does not want to use APS redundancy, then the 2 portcards can be installed in any two of portcard slots 1 through 4. If APS redundancy is desired, then the two portcards must be installed either in slots 1 and 2 or slots 3 and 4.

| Portcard Slot | Minimum # of Fabrics |
|---------------|----------------------|
| 1-4 | 1 |
| 5-8 | 2 |
| 9-12 | 3 |
| 13-16 | 4 |
| 17-24 | 6 |
| 25-48 | 12 |

Table 0: Fabric Requirements for Portcard Slot Usage

To add capacity, add the new fabric(s), wait for the switch to recognize the change and reconfigure the system to stripe across the new number of fabrics. Install the new portcards.

Note that it is not technically necessary to have the full 4 portcards per fabric. The switch will work properly with 3 fabrics installed and a single portcard in slot 12. This isn't cost efficient but it will work.

To remove capacity, reverse the adding capacity procedure.

If the switch is oversubscribed, i.e. install 8 portcards and only one fabric.

It should only come about as the result of improperly upgrading the switch or a system failure of some sort. The reality is that one of two things will occur, depending on how this situation arises. If the switch is configured as a 40G switch and the portcards are added before the fabric, then the 5th through 8th portcards will be dead. If the switch is configured as 80G non-redundant switch and the second fabric fails or is removed then all data through the switch will be corrupted (assuming the spare fabric is not installed). And just to be complete, if 8 portcards were installed in an 80G redundant switch and the second fabric failed or was removed, then the switch would continue to operate normally with the spare covering for the failed/removed fabric.

The switch includes the following features:

- Scales from 40Gbps to 480Gbps (40, 80, 120, 160, 240, 480 GB/sec are the supported configurations).
- Switches ATM cells and variable-length packets
- N+1 fabric redundancy with error detection and recovery supported in the ASIC chipset.
- Native APS support
- Support up to 196K cell shared memory, 9216K unicast and 64K multicast connections.
- Support 2x port speed for fabric dequeuing (2.5 GB/sec in, 5 GB/sec out for each OC48 port).

- Supports both OC48c ports and OC192c ports.
- Provides port/priority queuing similar to past switch fabrics. Four priorities are provided for 40-120 GB/sec switches, 2 priorities/port for 240 GB/sec switches and 1 priority for 480 GB/sec switches.
- ASICs utilize 250 MHz HSTL point to point busses between fabric ASICs and interface with the backplane using standard GBit transceivers.
- Interface to port cards chips use 80-125 MHz LVTTTL signals.
- Support output port supplied back-pressure.

The significant architectural difference between the switch and past switches is that incoming traffic is routed to multiple switch fabrics. Each fabric is designed to enqueue 40 GB/sec of data and dequeue 80 GB/sec of data. As data comes into the switch, it is broken up on a bit by bit basis and part of each packet is sent to each fabric in the box. The fabrics will all make the same enqueueing and drop decisions, and all schedule fragments of a packet/cell at the same time. Each fabric sends its portion of the packet or cell to the output port card which reassembles the fragment into the complete cell/packet which is then passed to a shared memory ASIC for per port storage and scheduling. The XOR of the data sent to each fabric is sent to a spare fabric. In the event of a fabric failure, that fabrics data can be recovered by utilizing the good data bits and the parity fabric bits to recalculate any fabrics data. The striping of data to fabrics

happens on the basis of 48 bit chunks. This allows the switch to support 1,2,3,4,6 and 12 fabrics.

Five ASICs build the switching functionality for the switch. These ASICs are described briefly below.

5 **TABLE 1:** The switch ASICs

| ASIC | Function |
|--------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Striper | Takes incoming cell from Vortex (or OC192c equivalent) or from POS input stage and breaks the data up into the appropriate chunks to go to each fabric, calculates the parity for the spare fabric, concatenates a checksum onto the packet, separates the routeword and data into separate routeword and data busses which run across the backplane. |
| Aggregator | Receives separate data and routeword busses from multiple stripers. Converts from the reasonably slim dedicated striper->Aggregator busses to a wide shared bus to the memory controllers. |
| Memory Controllers | Actually perform the queueing of data for the fabrics. Queues the cell into one of 200 queues (192 UC queues, 4 MC queues and 4 control port queues). All drops which occur in the chipset occur here. |
| Separator | Combines traffic from multiple memory controllers to one fabric output. Provides rate control of the stream of data leaving the fabric for each OC48 or OC192c port. |
| Unstriper | Receives data from multiple separators. Combines traffic and error checks the received data. Detects errors on any fabric and attempts to reconstruct the good data. Passes the data to the output memory controller. If the striper is on an ATM blade and the data is a packet, it is segmented before passing onto the ATM controller. |

Figure 1 shows packet striping in the switch.

The chipset supports ATM and POS port cards in both OC48 and OC192c configurations. OC48 port cards interface to the switching fabrics with four separate OC48 flows. OC192 port cards logically combine the 4 channels into a 10G stream. The ingress side of a port card does not perform traffic conversions for traffic changing between ATM cells and packets. Whichever form of traffic is received is sent to the switch fabrics. The switch fabrics will mix packets and cells and then dequeue a mix of packets and cells to the egress side of a port card.

The egress side of the port is responsible for converting the traffic to the appropriate format for the output port. This convention is referred to in the context of the switch as "receiver

makes right". A cell blade is responsible for segmentation of packets and a cell blade is responsible for reassembly of cells into packets. To support fabric speed-up, the egress side of the port card supports a link bandwidth equal to twice the inbound side of the port card. For each OC48 interface, the unstriper supports a bandwidth of 6GB/sec and for each OC192 interface, a bandwidth of 24 GB/sec (combined routeword + data).

The block diagram for a Poseidon-based ATM port card is shown as in Figure 2. Each 2.5G channel consists of 4 ASICs: Vortex and striper ASIC at the inbound side and unstriper ASIC and Trident ASIC at the outbound side.

At the inbound side, the Vortex ASIC aggregates 1 OC-48c or 4 OC-12c interfaces. Each vortex sends a 2.5G cell stream into a dedicated striper ASIC (using the BIB bus, as described below). The striper converts the vortex supplied routeword into two pieces. A portion of the routeword is passed to the fabric to determine the output port(s) for the cell. The entire routeword is also passed on the data portion of the bus as a routeword for use by the outbound memory controller. The first routeword is termed the "fabric routeword". The routeword for the outbound memory controller is the "egress routeword".

At the outbound side, the unstriper ASIC in each channel takes traffic from each of the port cards, error checks and correct the data and then sends correct packets out on its output bus. The unstriper uses the data from the spare fabric and the checksum inserted by the striper to detect and correct data corruption. The 5Gbps traffic is then sent to the Trident ASIC of the Poseidon chipset. The Trident ASIC stores the incoming cells based on per-VC

queues and sends them out to OC- 12c/OC-48c interfaces at aggregated speed of 2.5Gbps.

For the POS interfaces, the striper ASIC input bus speeds up to 3.2Gbps to handle POS overhead. The outbound side, the
5 unstriper talks to a reassembly stage which is currently being defined.

Figure 2 shows an OC48 Port Card.

The OC192 port card supports a single 10G stream to the fabric and between a 10G and 20G egress stream. This board also
10 uses 4 stripers and 4 unstriper, but the 4 chips operate in parallel on a wider data bus. The data sent to each fabric is identical for both OC48 and OC192 ports so data can flow between the port types without needing special conversion functions.

Figure 3 shows a 10G concatenated network blade.

Each 40G switch fabric enqueues up to 40Gbps cells/frames and dequeue them at 80Gbps. This 2X speed-up reduces the amount of traffic buffered at the fabric and lets the outbound ASIC digest bursts of traffic well above line rate. A switch fabric consists of three kinds of ASICs: aggregators, memory controllers, and
20 separators. Nine aggregator ASICs receive 40Gbps of traffic from up to 48 network blades and the control port. The aggregator ASICs combine the fabric route word and payload into a single data stream and TDM between its sources and places the resulting data on a wide output bus. An additional control bus (destid) is used to control
25 how the memory controllers enqueue the data. The data stream from each aggregator ASIC then bit sliced into 12 memory controllers.

The memory controller receives up to 16 cells/frames every 250MHz clock cycle. Each of 12 ASICs stores 1/12 of the aggregated data streams. It then stores the incoming data based on control information received on the destid bus. Storage of data is simplified in the memory controller to be relatively unaware of packet boundaries (cache line concept). All 12 ASICs dequeue the stored cells simultaneously at aggregated speed of 80Gbps.

90 Nine separator ASICs perform the reverse function of the aggregator ASICs. Each separator receives data from all 12 memory controllers and decodes the routewords embedded in the data streams by the aggregator to find packet boundaries. Each separator ASIC then sends the data to up to 24 different unstripers depending on the exact destination indicated by the memory controller as data was being passed to the separator.

95 The dequeue process is back-pressure driven. If back-pressure is applied to the unstriper, that back-pressure is communicated back to the separator. The separator and memory controllers also have a back-pressure mechanism which controls when a memory controller can dequeue traffic to an output port.

20 In order to support OC48 and OC192 efficiently in the chipset, the 4 OC48 ports from one port card are always routed to the same aggregator and from the same separator (the port connections for the aggregator & Sep are always symmetric.). The table below shows the port connections for the aggregator & sep on each fabric for the switch configurations. Since each aggregator is accepting traffic from 10G of ports, the addition of 40G of switch capacity only adds ports to 4 aggregators. This leads to a differing port connection pattern for the first four aggregators from the second 4 (and also the corresponding separators).

TABLE 2: Agg/Sep port connections

| Switch Size | Agg 1 | Agg 2 | Agg 3 | Agg 4 | Agg 5 | Agg 6 | Agg 7 | Agg 8 |
|-------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| 40 | 1,2,3,4 | 5,6,7,8 | 9,10,11,12 | 13,14,15, 16 | | | | |
| 80 | 1,2,3,4 | 5,6,7,8 | 9,10,11,12 | 13,14,15, 16 | 17,18,19, 20 | 21,22,23, 24 | 25,26,27, 28 | 29,30,31, 32 |
| 120 | 1,2,3,4 | 5,6,7,8 | 9,10,11,12 | 13,14,15, 16 | 17,18,19, 20 | 21,22,23, 24 | 25,26,27, 28 | 29,30,31, 32 |
| 160 | 33,34,35, 36 | 37,38,39, 40 | 41,42,43, 44 | 45,46,47, 48 | | | | |
| | 1,2,3,4 | 5,6,7,8 | 9,10,11,12 | 13,14,15, 16 | 17,18,19, 20 | 21,22,23, 24 | 25,26,27, 28 | 29,30,31, 32 |
| | 33,34,35, 36 | 37,38,39, 40 | 41,42,43, 44 | 45,46,47, 48 | 49,50,51, 52 | 53,54,55, 56 | 57,58,59, 60 | 61,62,63, 64 |

Figure 4 shows the connectivity of the fabric ASICs.

The external interfaces of the switches are the Input Bus (BIB) between the striper ASIC and the ingress blade ASIC such as Vortex and the Output Bus (BOB) between the unstriper ASIC and the egress blade ASIC such as Trident.

Two variations of routewords are supported. The first option uses one 32 bit routeword which is passed to the egress board as the egress routeword and has fields extracted to form the fabric routeword. The second option allows the striper to accept both a fabric routeword (which happens on a dedicated routeword bus) and an egress routeword (which is received on the data bus). The second option is more flexible on connection space usage and expansion since that allows all 32 bits of the routeword to be used to identify connections on switch egress.

To maintain compatibility with Vortex, bit 24 is still maintained as the multicast bit. The incoming routeword has the following format.

| | | |
|------------------------------------------------|---------------|----------------------------------------------|
| bit 30:25 | bit 24 | bit 23:0 |
| Connection ID(29:28) & Connection ID(19:16) | Multicast Bit | Connection ID (27:20) & connection ID (15:0) |

For UC connections, although bits 29:5 are passed to the fabric, only bits 29:20 are used. These bits should be programmed with queue to be used. Bits 29:28 should be programmed with the priority and bits 27:20 programmed with the queue number.

If the fabric is using 10 bits of conn ID, this leaves 20 bits (1 M connections) for use by the outbound memory controller.

For double routewords, no manipulation is done. The value passed in on the routeword bus needs to equal to the connection ID to be transmitted on the backplane. The following two tables show the routeword value which should be passed on the

25 backplane routeword bus.

TABLE 4: Unicast Connection ID for separate RW bus

| bit 25 | bit 24:23 | Bit 22:15 | 14:0 |
|-----------------|-----------------|-----------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|
| Multicast bit=0 | Fabric priority | Fabric queue ID | Future expansion bits. This bits are transmitted to the fabric, but the current fabric ignores them. Future fabrics may expand to support these bits. |

TABLE 5: Multicast Connection ID for separate RW bus

| bit 25 | bit 24:23 | bit 22:16 | bit 15:0 |
|-----------------|-------------------|------------------------------------------------------------------------------------------------------------|--------------------------------------------------------|
| Multicast bit=1 | Priority queue ID | Reserved. Note these bits are sent to the fabric to allow future fabrics to support more connection space. | multicast connection ID (0 to 64K) used by the fabric. |

Special routewords are flagged by using reserved queue numbers (those in the range of 248-255). These routeword values indicate the receipt of an OAM cell which must get routed to the control port or a queue resynch operation. These special values are always expressed in terms of the connection ID which goes to the fabric. If special routewords are given to the fabric, the memory controller routeword must also be modified if these are getting passed in using the separate connection number bus.

The routeword passed to the fabric will contain the multicast bit and the port mask bits (bits 23:16). The routeword passed to the outbound memory controller will maintain the port mask and also contain the vortex ID and the port ID.

The connection ID of an OAM cell has a special format generated by the Vortex ASIC:

TABLE 6: Connection ID for OAM cell

| bit 25 | Bit 24:23 | bit 22:15 | bit 14:9 | 8 | bit 7:0 |
|-----------------|-----------------|------------|-----------------|----------|---------|
| Multicast bit=0 | Vortex ID (7:6) | 0xF0 (hex) | Vortex ID (5:0) | reserved | Port ID |

The Vortex ID field is used to indicate which source Vortex ASIC the cell comes from. The port ID indicates which port the cell comes from inside the Vortex ASIC. Note that OAM cells are all unicast. All OAM cells are destined to one of 196 blade and control port queues programmed by a 8-bit OAM cell destination register in the memory controller ASICs. If separate routeword busses are being used, bit 24:16 of the BIB_CONN field will be passed to the fabric. The routeword which appears on the data bus (memory controller routeword) should include the port mask, vortex ID and port ID fields in bits 23:0. The value in the multicast bit is a don't care for the memory controller routeword.

Fabric queue ID 0xF0-0xF7 of the unicast connection ID is reserved for software use. All packets which have the fabric queue ID in range of 0xF0-0xFF will be redirected to one of the 4 control port queues based on a programmable register.

The connection ID of a resync cell has the following format. The resync cell is used to resynchronize queues in the memory controller ASICs. Fabric queue ID 0xF8-0xFF of the unicast connection ID is reserved for special fabric functions.

TABLE 7: Connection ID for Resync cell

| bit 25 | 24:23 | bit 22:15 | bit 14:13 | bit 12:0 |
|-----------------|-------------------|------------|-------------------------------|----------|
| Multicast bit=0 | Priority (unused) | 0xFF (hex) | Number of priorities per port | Reserved |

The number of priority queues per port can only be changed during the queue resync period, i.e., when a fabric is removed or inserted as follows:

- 00: one priority per port for 480G switch, pick bit 15 down to 8 of the connection ID as the queue ID;
- 01: two priorities per port for 240G switch, pick bit 16 down to 9 of the connection ID as the queue ID;
- 10: 4 priorities per port for 120G or smaller switch, pick bit 17 down to 10 of the connection ID as the queue ID;
- 11: reserved

The resync cell can also be used to copy the shadow data register to a valid location where the shadow address register points to.

Shadow control cell is used to copy the shadow data register to a valid location where the shadow address register points to. The connection ID of a shadow control cell use.

TABLE 8: Connection ID for Shadow Control Cell

| bit 24 | 24:23 | bit 22:15 | bit 14:0 |
|-----------------|----------|-----------|----------|
| Multicast bit=0 | Priority | 0xFE(hex) | Reserved |

Data coming into the BIB bus and out of the BOB bus is assumed to be filled onto the busses from most significant bit to least significant bit (highest number bit to lowest number bit).

The Striper ASIC accepts data from the ingress port via the Input Bus (BIB) (also known as DIN_ST_bl_ch bus).

This bus can either operate as 4 separate 32 bit input buses (4xOC48c) or a single 128 bit wide data bus with a common set of control lines to all stripers. This bus supports either cells or packets based on software configuration of the striper chip. It consists of the following signals:

- BIB_Clock: This clock is sourced by the Striper ASIC at up to 100 MHz and is used as a reference for data and control signals on the BIB.
- BIB_BP: This signal is asserted (low) to indicate the striper ASIC cannot take data on the bus due to a bandwidth difference between the BIB and SIB busses. Interfaces which run below 93 MHz will never see this signal asserted. At 100 Mhz, this signal is asserted if more than 65536 bytes of back-to-back data are given. This signal should be sampled at the start of packet. During a packet transfer, this signal will be asserted if the FIFO conditions would cause BP if the packet ended on the current clock cycle. If BP is asserted the clock cycle after the EOP, the striper will effectively ignore the input bus until the BP indication is withdrawn. The packet ingress stage should repeat the first word of the next packet transfer and then proceed with the rest of the packet after the BP signal goes away.
- BIB_Valid_L: This active low input signal delimits valid data on the BIB_SOP, BIB_EOP, and BIB_DATA busses. If this signal is active, the busses are assumed to be valid. If high, the busses are treated as having invalid data for the current clock cycle. If a transfer is not in progress (no SOP without EOP has been given) then the

data bus is treated as invalid even if this signal is a one. For cell interfaces, this signal can be tied active.

- BIB_Cell_Pkt: This signal is set to a one to indicate a cell transfer and a zero to indicate a packet transfer. Signal needs to be valid the same clock cycle as start of cell.

- BIB_Data[127:0]: This is the input 128-bit data bus. If running in 32 bit mode, a cell consists of a 4 byte RW, a 4 byte Header, and twelve 4 byte data words. A packet has a RW and N data words, where $1 \leq N$. If running in 128 bit mode, a cell has a 4 byte RW, a 4 byte header, and 8 bytes of data in the first word, 2 words with 16 bytes of data, and a final word with 8 bytes of data, if the data starts on a word boundary. A following cell can start on the half-word boundary and have all fields offset by 8 bytes. Packets in 128 bit mode work in the same fashion as 32 bit mode, except that EOP and SOP can have larger values. Minimum packet length supported is 16 bytes. If half-word boundary cell starts are used, the correct value (0/4) needs to be given on the SOP bits 3:0.

- BIB_EOP[4:0]: This bus has two fields. Bit 4 is a one to indicate an EOP on the current transfer (if BIB_Valid_L is active). Bit 4 is a zero to indicate no EOP on the current transfer. Bits 3:0 give the offset of the last byte which is valid. The EOP field is not utilized for cell transfers.

- BIB_SOP/C[1:0]: This bit indicates a start of packet or cell on the current bus cycle (if BIB_Valid_L is active).

A value of zero indicates start of transfer, a value of one indicates no start of transfer. Asserting bit 1 = 1 indicates that the upper 64 bits carries the SOP and asserting bit 0=1 indicates that the lower 64 bits carries the SOP (for 128 bit bus only). For the 32 bit bus, SOP(0) should be used, SOP(1) should be tied high. For the 128 bit bus, if a packet ends in the upper 64 bits of the bus, a new packet can begin at bit 64.

- BIB_CONN(24:0): This is an optional bus. It can be used to pass a routeword to the striper ASIC to use as the fabric routeword, or the routeword can be transferred as the most significant 32 bits of the first word of data. The data should be valid the same cycle as SOP/C. The value during non-SOP/C cycles is a don't care. The interface is statically configured to either use the separate connection number bus or to expect the routeword on the data bus.

Figure 5 shows a 32 bit BIB cell transfer.

Figure 6 shows a BIB back-pressure.

Figure 7 shows a 32 bit BIB packet transfer using external connection number bus.

The unstriper ASIC sends data to the egress port via Output Bus (BOB) (also known as DOUT_UN_b1_ch bus), which is a 64 (or 256) bit data bus that can support either cell or packet. It consists of the following signals:

This bus can either operate as 4 separate 32 bit output buses (4xOC48c) or a single 128 bit wide data bus with a common set of control lines from all Unstripers. This bus supports either cells or packets based on software configuration of the unstriper chip. It consists of the following signals:

- BOB_Clock: This clock is sourced from the unstriper ASIC at up to 100 MHz and is used as a reference for data and control signals on the BOB.
- BOB_BP: This active low input signal indicates whether data can be transferred (inactive) or cannot be transferred (active). When back-pressure is asserted, the unstriper will stop advancing the output bus and signal data is not valid using the BOB_valid signal. Since synchronization must be done on both sides of the interfaces, 8 clock cycles of data must be allowed from the assertion of BP to data stopping. The source driving BOB_BP cannot make any assumptions on the data stopping or restarting except by examining BOB_Valid.
- BOB_Valid_L: This active low output signal indicates whether the bus has valid data or not during a transfer. This signal indicates invalid data only when BOB_BP has been asserted.
- BOB_Data: This is the output bit data bus. It can either be 64 bits wide or 256 bits wide. If running in 64 bit mode, a cell consists of a word with a 4 byte RW and a 4 byte Header followed by 6 data words. A packet has a RW and N data words, where $1 \leq N$. If running in 256 bit mode and a cell starts on an even 32 byte word boundary, a

cell has a word with a 4 byte RW a 4 byte header and 24 bytes of data in the first word, and a second word with 24 bytes of data. A following cell can start on the next used byte and have all fields offset by 8 bytes. Valid cell start locations are all multiples of 8 (0, 8, 16, 24). Packets in 128 bit mode work in the same fashion as 32 bit mode, except that EOP and SOP can have larger values. Minimum packet length supported is 16 bytes. If half-word boundary cell starts are used, the correct value (0/4) needs to be given on the SOP bits 3:0.

- BOB_EOP: This bit is asserted when the last transfer of a packet is occurring.
- BOB_Cell_Pkt: This signal is set to a one to indicate a cell transfer and a zero to indicate a packet transfer. Signal needs to be valid the same clock cycle as start of cell.
- BOB_SOP/C This bit is a zero to indicate a start of packet or cell on the current bus cycle. Data is always assumed to start at the most significant bit of the bus.

Figure 8 shows a 64 bit BOB cell transfer.

Figure 9 shows a 64 bit BOB packet transfer.

Figure 10 shows an overview of the datapath of the switch ASICs.

The data on the data bus transports an optional byte count (32 bit word, lower 16 bits are the byte count) and a 32 bit egress routeword. The unstriper core will always produce a byte count. If a segmentation engine is used to break the packet up into cells, then the segmentation engine will drop the byte count word before it is given to the cell interface. This dropping is only supported in OC48 mode. In OC192 mode, the chipset will have no provisions for segmentation and dropping the byte count word.

TABLE 9: OC48 BOB format

| OC48 Bits | OC192 bits | Label | Usage |
|-----------|------------|------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 63:48 | 255:240 | Unused | reserved for unstriper use |
| 47:32 | 239:224 | Byte count | Gives the count of the number of bytes in the packet not counting the 4 bytes for the egress routeword and the bytes for the byte count (basically, this corresponds to the byte count of the received packet plus/minus any changes for reencapsulation, pushes, or pops.) |
| 31:0 | 223:192 | Egress RW | Routeword for the egress memory controller Next bits start the data (bits (191 to 0) for 192, next clock cycle for OC48 |

The Synchronizer has two main purposes. The first purpose is to maintain logical cell/packet or datagram ordering across all fabrics. On the fabric ingress interface, datagrams arriving at more than one fabric from one port cards's channels need to be processed in the same order across all fabrics. The Synchronizer's second purpose is to have a port cards's egress channel re-assemble all segments or stripes of a datagram that belong together even though the datagram segments are being sent from more than one fabric and can arrive at the blade's egress inputs at different times. This mechanism needs to be maintained in a system that will have different net delays and varying amounts of clock drift between blades and fabrics.

The switch uses a system of a synchronized windows where start information is transmit around the system. Each transmitter

and receiver can look at relative clock counts from the last resynch indication to synchronize data from multiple sources. The receiver will delay the receipt of data which is the first clock cycle of data in a synch period until a programmable delay after it receives the global synch indication. At this point, all data is considered to have been received simultaneously and fixed ordering is applied. Even though the delays for packet 0 and cell 0 caused them to be seen at the receivers in different orders due to delays through the box, the resulting ordering of both streams at receive time = 1 is the same, Packet 0, Cell 0 based on the physical bus from which they were received.

Multiple cells or packets can be sent in one counter tick. All destinations will order all cells from the first interface before moving onto the second interface and so on. This cell synchronization technique is used on all cell interfaces. Differing resolutions are required on some interfaces.

The Synchronizer consists of two main blocks, mainly, the transmitter and receiver. The transmitter block will reside in the Striper and Separator ASICs and the receiver block will reside in the Aggregator and Unstripers ASICs. The receiver in the Aggregator will handle up to 24(6 port cards x 4 channels) input lanes. The receiver in the Unstripers will handle up to 13(12 fabrics + 1 parity fabric) input lanes.

When a sync pulse is received, the transmitter first calculates the number of clock cycles it is fast (denoted as N clocks).

The transmit synchronizer will interrupt the output stream and transmit N K characters indicating it is locking down.

At the end of the lockdown sequence, the transmitter transmits a K character indicating that valid data will start on the next clock cycle. This next cycle valid indication is used by the receivers to synchronize traffic from all sources. Refer to "K character usage" on page 34 for the mapping of K characters to the functions.

At the next end of transfer, the transmitter will then insert at least one idle on the interface. These idles allow the 10 bit decoders to correctly resynchronize to the 10 bit serial code window if they fall out of synch.

The receive synchronizer receives the global synch pulse and delays the synch pulse by a programmed number (which is programmed based on the maximum amount of transport delay a physical box can have). After delaying the synch pulse, the receiver will then consider the clock cycle immediately after the synch character to be eligible to be received. Data is then received every clock cycle until the next synch character is seen on the input stream. This data is not considered to be eligible for receipt until the delayed global synch pulse is seen.

Since transmitters and receivers will be on different physical boards and clocked by different oscillators, clock speed differences will exist between them. To bound the number of clock cycles between different transmitters and receivers, a global sync pulse is used at the system level to resynchronize all sequence counters. Each chip is programmed to ensure that under all valid clock skews, each transmitter and receiver will think that it is fast by at least one clock cycle. Each chip then waits for the appropriate number of clock cycles they are into their current sync_pulse_window. This ensure that all sources run N* sync_pulse_window valid clock cycles between synch pulses.

As an example, the synch pulse window could be programmed to 100 clocks, and the synch pulses sent out at a nominal rate of a synch pulse every 10,000 clocks. Based on a worst case drifts for both the synch pulse transmitter clocks and the synch pulse receiver clocks, there may actually be 9,995 to 10,005 clocks at the receiver for 10,000 clocks on the synch pulse transmitter. In this case, the synch pulse transmitter would be programmed to send out synch pulses every 10,006 clock cycles. The 10,006 clocks guarantees that all receivers must be in their next window. A receiver with a fast clock may have actually seen 10,012 clocks if the synch pulse transmitter has a slow clock. Since the synch pulse was received 12 clock cycles into the synch pulse window, the chip would delay for 12 clock cycles. Another receiver could see 10,006 clocks and lock down for 6 clock cycles at the end of the synch pulse window. In both cases, each source ran 10,100 clock cycles.

When a port card or fabric is not present or has just been inserted and either of them is supposed to be driving the inputs of a receive synchronizer, the writing of data to the particular input FIFO will be inhibited since the input clock will not be present or unstable and the status of the data lines will be unknown. When the port card or fabric is inserted, software must come in and enable the input to the byte lane to allow data from that source to be enabled. Writes to the input FIFO will be enabled. It is assumed that, the enable signal will be asserted after the data, routeword and clock from the port card or fabric are stable.

At a system level, there will be a primary and secondary sync pulse transmitter residing on two separate fabrics. There will also be a sync pulse receiver on each fabric and blade. This

can be seen in Figure 11. A primary sync pulse transmitters will be a free-running sync pulse generator and a secondary sync pulse transmitter will synchronize its sync pulse to the primary. The sync pulse receivers will receive both primary and secondary sync pulses and based on an error checking algorithm, will select the correct sync pulse to forward on to the ASICs residing on that board. The sync pulse receiver will guarantee that a sync pulse is only forwarded to the rest of the board if the sync pulse from the sync pulse transmitters falls within its own sequence "0" count. For example, the sync pulse receiver and an Unstriper ASIC will both reside on the same Blade. The sync pulse receiver and the receive synchronizer in the Unstriper will be clocked from the same crystal oscillator, so no clock drift should be present between the clocks used to increment the internal sequence counters. The receive synchronizer will require that the sync pulse it receives will always reside in the "0" count window.

If the sync pulse receiver determines that the primary sync pulse transmitter is out of sync, it will switch over to the secondary sync pulse transmitter source. The secondary sync pulse transmitter will also determine that the primary sync pulse transmitter is out of sync and will start generating its own sync pulse independently of the primary sync pulse transmitter. This is the secondary sync pulse transmitter's primary mode of operation. If the sync pulse receiver determines that the primary sync pulse transmitter has become in sync once again, it will switch to the primary side. The secondary sync pulse transmitter will also determine that the primary sync pulse transmitter has become in sync once again and will switch back to a secondary mode. In the secondary mode, it will sync up its own sync pulse to the primary sync pulse. The sync pulse receiver will have less tolerance in its sync pulse filtering mechanism than the secondary sync pulse

[illegible]

01

15

20

25

Lockdown may occur at different times on different chips. All fabric input FIFOs are initially set up such that lockdown can occur on either side of the FIFO first without the FIFO running dry or overflowing. On each chip-chip interface, there is a sync FIFO
5 to account for lockdown cycles (as well as board trace lengths and clock skews). The transmitter signals lockdown while it is locked down. The receiver does not push during indicated cycles, and does not pop during its own lockdown. The FIFO depth will vary depending on which chip locks down first, but the variation is
10 bounded by the maximum number of lockdown cycles. The number of lockdown cycles a particular chip sees during one global sync period may vary, but they will all have the same number of useful cycles. The total number of lockdown cycles each chip on a particular fabric sees will be the same, within a bounded
15 tolerance.

The Aggregator core clock domain completely stops for the lockdown duration - all flops and memory hold their state. Input FIFOs are allowed to build up. Lockdown bus cycles are inserted in the output queues. Exactly when the core lockdown is executed is
20 dictated by when DOUT_AG bus protocol allows lockdown cycles to be inserted. DOUT_AG lockdown cycles are indicated on the DestID bus.

The memory controller must lockdown all flops for the appropriate number of cycles. To reduce impact to the silicon area in the memory controller, a technique called propagated lockdown is
25 used.

The aggregator signals lockdown cycles on the DIN_ME bus. The memory controller does not push during these cycles. The memory controller does not pop during lockdown to account for the non-push cycles. The FIFO depth is set during fabric

synchronization to tolerate getting deeper or shallower depending on who locks down first.

Lockdown idle cycles are inserted on the DOUT and CH_ID busses. An extended sync signal is used to indicate the number of lockdown cycles on the DOUT_ME bus to aid the Separator's lockdown function.

03010 The token bus lockdown looks the same as the DIN_ME bus from a memory controller perspective. Non-push cycles are signaled by the separators according to their lockdowns. The memory controller does not pop during lockdown. The Separator locks down completely in a manner similar to the Aggregator. DIN_SP and CH_ID lockdown cycles are signaled individually per-bus via the SYNC signals. Any continuous SYNC assertion after the first one is considered a lockdown cycle. Lockdown bus cycles are not pushed into the input FIFOs.

03020 The chip-to-chip communication within a single fabric must be synchronized. Although no clock drift exists between chips, differences in track delays cause data to arrive at different Memory Controllers at different times. All Memory Controllers need to process incoming packets in exactly the same logical order on each chip. The Separators must align and combine multiple data slices coming from different Memory Controllers. The Memory Controllers must take the tokens received from the Separators and apply them at exactly the same point in the logical packet flow, or drop decisions may differ from chip to chip.

The on-fabric chip-to-chip synchronization is executed at every sync pulse. While some sync error detecting capability may exist in some of the ASICs, it is the Unstriper's job to detect

fabric synchronization errors and to remove the offending fabric. The chip-to-chip synchronization is a cascaded function that is done before any packet flow is enabled on the fabric. The synchronization flows from the Aggregator to the Memory Controller,
5 to the Separator, and back to the Memory Controller. After the system reset, the Aggregators wait for the first global sync signal. When received, each Aggregator transmits a local sync command (value 0x2) on the DestID bus to each Memory Controller.

The Memory Controllers do not push anything into a DIN
10 input FIFO until the first sync command is seen on that bus. The sync and every bus cycle following is constantly pushed into the input FIFO. On the core side of the input FIFOs, no FIFO is popped until a sync appears in the FIFO from every Aggregator. After two additional margin cycles, every input FIFO is popped every cycle.
15 After this point the input FIFO depths remain constant. The depths are roughly a function of the track delays from each Aggregator. Immediately after the Memory Controllers begin sampling the Aggregator input FIFOs, a sync signal (S_SYNC_L) is transmitted to all Separators on the DOUT and CH_ID busses.

20 Like the Memory Controllers, the Separators do not push into the DIN and CH_ID busses until a sync signal is received on that bus. The sync and everything after is constantly pushed into the input FIFO.

On the core side the Separator always waits until at
25 least one word is present on all input busses, and then pops the CH_ID and DIN busses simultaneously. This will logically align the data stripes coming from the Memory Controllers. After the first combined sync is popped from the input FIFOs, the Separators send a sync signal on the TOKEN bus to the Memory Controllers.

The Memory Controllers do not push into the TOKEN bus input FIFO until a sync signal (0x3F on the token bus) has been seen on the bus. The sync and all subsequent tokens and idles are always pushed.

5 All Memory Controllers need to apply the received tokens to the same point in the incoming logical flow in order for all drop decisions to be identical. This is done by waiting a worst case number of clock cycles after the Separator sync transmission before beginning to pop the token input FIFO. The worst case delay
10 must be used because there is no way for a single Memory Controller to know exactly when all other Memory Controllers have received a token. The programmable delay stored in the 16-bit Token Sync Wait Register is in "useful" cycles (125MHz) that do not include the fabric lockdown cycles. The worst case delay is the worst case
15 skew for all data paths going from the Aggregator to Memory Controller to Separator and back to Memory Controller.

The following Table 11 gives the min/max delays which the chipset supports and represent the limits of what is verified in the chip verification process.

20 Sync pulse transport delay from Transmitter to any individual chip receiving the sync pulse(WC path - BC path): 500 nS (min delay of 0, max delay of 500 nS). At 175 ps/inch, this works out to a difference of about 70m. Backplane transport delay difference from local sync pulse receipt to reception of the sync
25 indication flag by the far end chips: 500 nS. Note that it is desired to allot about 25 ns of this to the chip synchronizer operation which gives a delta path delay supported of 500 nS.

Oscillators should be 100 ppm oscillators. The assumption of the design was that the difference in transmission path delay was less than or equal to clock drift. On board delays between chips have been designed to exceed the following specs:

- 5 Shortest net: 0.25", transport delay of pretty much 0.
Longest net: 25", transport delay is 5 nS.

For any signal distribution. The net delta delay between chips is a multiplier of the number of busses the sync has traversed. Since the sync goes through a receive synchronization to the local clock of the chip, an +/- 8 nS uncertainly has to be added at each stage giving a net uncertainty of around 21 nS for each hop.

TABLE 11: Fabric sync delay

| Chip | Number of busses | Skew | Notes |
|----------------------------|------------------|-------|-------------------------------------------------------------------------------------------------------------------------|
| Agg | 1 | 21 nS | Sync pulse in |
| Memory controller DIN | 2 | 42 nS | Sync pulse to agg + agg_mc delta |
| Sep DIN | 3 | 63 nS | Sync pulse to agg + agg_mc + mc_sep (note this sync pulse is delayed by the memory controller for propagated lockdown). |
| memory controller token_in | 4 | 84nS | Everything above + sep_mc tokens. |

The control port follows the same cell flow as the regular ports. The switch control processor sends cells to the striper ASIC; the striper stripes the cells and route words across all fabrics. An additional aggregator (9th) ASIC sends cells via the DOUT_AG/DestID buses to all 12 memory controllers. Each memory controller ASIC has an additional 9th DIN_ME_fb_se_9 bus.

The memory controller ASIC will route the incoming control port cells to any one of the control port destination queues and blade queues (up to 196 queues). The 9th DOUT_ME_fb_se_9 bus is used to send the control cells to the 9th separator ASIC, which sends the cells to one of several destination unstriper ASICs. The unstriper ASIC reconstructs the cells from all 9th separator ASICs across all fabrics. It sends the complete control cells to the switch control processor it is connected to.

Note that the control port destination queues can be part of any multicast cells such that the multicast port mask is necessary to include additional bit(s) to indicate the control port queue(s).

There are at most 4 control ports in any switch configurations. This limitation is due to the aggregator and separator ASICs only have 4 12-bit channels which can be scalable to different switch configurations, respectively. In other words, bus DIN_AG_fb_9_1_1, DIN_AG_fb_9_2_1, DIN_AG_fb_9_3_1, and DIN_AG_fb_9_4_1 of the aggregator ASIC are connected to up to 4 control port striper ASICs. Bus DOUT_SP_fb_9_1_1, DOUT_SP_fb_9_2_1, DOUT_SP_fb_9_3_1, and DOUT_SP_fb_9_4_1 of the separator ASIC are connected to up to 4 control port unstriper ASICs.

The striping function assigns bits from incoming data streams to individual fabrics. Two items were optimized in deriving the striping assignment:

1. Backplane efficiency should be optimized for OC48 and OC192
2. Backplane interconnection should not be significantly altered for OC192 operation.

These were traded off against additional muxing legs for the striper and unstriper ASICs. Irregardless of the optimization, the switch must have the same data format in the memory controller for both OC48 and OC192.

5 Backplane efficiency requires that minimal padding be added when forming the backplane busses. Given the 12 bit backplane bus for OC48 and the 48 bit backplane bus for OC192, an optimal assignment requires that the number of unused bits for a transfer to be equal to $(\text{number_of_bytes} * 8) / \text{bus_width}$ where "/" is integer
0 division. For OC48, the bus can have 0, 4 or 8 unutilized bits. For OC192 the bus can have 0, 8, 16, 24, 32, or 40 unutilized bits.

This means that no bit can shift between 12 bit boundaries or else OC48 padding will not be optimal for certain packet lengths.

5 For OC192c, maximum bandwidth utilization means that each striper must receive the same number of bits (which implies bit interleaving into the stripers). When combined with the same backplane interconnection, this implies that in OC192c, each stripe must have exactly the correct number of bits come from each striper
20 which has 1/4 of the bits.

For the purpose of assigning data bits to fabrics, a 48 bit frame is used. Inside the striper is a FIFO which is written 32 bits wide at 80-100 MHz and read 24 bits wide at 125 MHz. Three 32 bit words will yield four 24 bit words. Each pair of 24 bit words
25 is treated as a 48 bit frame. The assignments between bits and fabrics depends on the number of fabrics.

TABLE 12: Bit striping function

| | | Fab 0 | Fab 1 | Fab 2 | Fab 3 | Fab 4 | Fab 5 | Fab 6 | Fab 7 | Fab 8 | Fab 9 | Fab 10 | Fab 11 |
|--------|-------|----------------------------|------------------------|----------------|------------|--------|--------|-------|-------|-------|-------|--------|--------|
| | 0:11 | 0:11 | | | | | | | | | | | |
| 1 fab | 12:23 | 12:23 | | | | | | | | | | | |
| | 24:35 | 24:35 | | | | | | | | | | | |
| | 36:47 | 36:47 | | | | | | | | | | | |
| | 0:11 | 0, 2, 5, 7, 8, 10 | 1, 3, 4, 6, 9, 11 | | | | | | | | | | |
| 2 fab | 12:23 | 13, 15, 16, 18, 21, 20, 22 | 12, 14, 17, 19, 20, 22 | | | | | | | | | | |
| | 24:35 | +24 to 0:11 | +24 to 0:11 | | | | | | | | | | |
| | 36:47 | +24 to 12:23 | +24 to 12:23 | | | | | | | | | | |
| | 0:11 | 0, 3, 5, 10 | 2, 4, 7, 9 | 1, 6, 8, 11 | | | | | | | | | |
| 3 fab | 12:23 | 15, 17, 22, 13 | 14, 16, 19, 21 | 13, 18, 20, 23 | | | | | | | | | |
| | 24:35 | +24 to 0:11 | +24 to 0:11 | +24 to 0:11 | | | | | | | | | |
| | 36:47 | +24 to 12:23 | +24 to 12:23 | +24 to 12:23 | | | | | | | | | |
| | 0:11 | 0, 5, 10 | 3, 4, 9 | 2, 7, 8 | 1, 6, 11 | | | | | | | | |
| 4fab | 12:23 | 15, 16, 21 | 14, 19, 20 | 13, 18, 23 | 12, 17, 22 | | | | | | | | |
| | 24:35 | 26, 31, 32 | 25, 30, 35 | 24, 29, 34 | 27, 28, 33 | | | | | | | | |
| | 36:47 | 37, 42, 47 | 36, 41, 46 | 39, 40, 43 | 38, 43, 44 | | | | | | | | |
| | 0:11 | 0, 11 | 1, 4 | 5, 8 | 2, 9 | 3, 6 | 7, 10 | | | | | | |
| 6 fab | 12:23 | 14, 21 | 15, 18 | 19, 22 | 12, 23 | 13, 16 | 17, 20 | | | | | | |
| | 24:35 | +24 to 0:11 | | | | | | | | | | | |
| | 36:47 | +24 to 12:23 | | | | | | | | | | | |
| | 0:11 | 0 | 4 | 8 | 1 | 5 | 9 | 2 | 6 | 10 | 3 | 7 | 11 |
| 12 fab | 12:23 | 15 | 19 | 23 | 12 | 16 | 20 | 13 | 17 | 21 | 14 | 18 | 22 |
| | 24:35 | 26 | 30 | 34 | 27 | 31 | 35 | 24 | 28 | 32 | 25 | 29 | 33 |
| | 36:47 | 37 | 41 | 45 | 38 | 42 | 46 | 39 | 43 | 47 | 37 | 40 | 44 |

The following tables give the byte lanes which are read first in the aggregator and written to first in the separator. The four channels are notated A,B,C,D. The different fabrics have

different read/write order of the channels to allow for all busses to be fully utilized.

One fabric-40G

5 The next table gives the interface read order for the aggregator.

| Fabric | 1st | 2nd | 3rd | 4th |
|--------|-----|-----|-----|-----|
| 0 | A | B | C | D |
| Par | A | B | C | D |

Two fabric-80G

| Fabric | 1st | 2nd | 3rd | 4th |
|--------|-----|-----|-----|-----|
| 0 | A | C | B | D |
| 1 | B | D | A | C |
| Par | A | C | B | D |

120G

| Fabric | 1st | 2nd | 3rd | 4th |
|--------|-----|-----|-----|-----|
| 0 | A | D | B | C |
| 1 | C | A | D | B |
| 2 | B | C | A | D |
| Par | A | D | B | C |

Three fabric-160G

| Fabric | 1st | 2nd | 3rd | 4th |
|--------|-----|-----|-----|-----|
| 0 | A | B | C | D |
| 1 | D | A | B | C |
| 2 | C | D | A | B |
| 3 | B | C | D | A |
| Par | A | B | C | D |

Siz fabric-240 G

| Fabric | 1st | 2nd | 3rd | 4th |
|--------|-----|-----|-----|-----|
| 0 | A | D | C | B |
| 1 | B | A | D | C |
| 2 | B | A | D | C |
| 3 | C | B | A | D |
| 4 | D | C | B | A |
| 5 | D | C | B | A |
| Par | A | C | D | B |

Twelve Fabric-480 G

| Fabric | 1st | 2nd | 3rd | 4th |
|---------|-----|-----|-----|-----|
| 0,1,2 | A | D | C | B |
| 3,4,5 | B | A | D | C |
| 6,7,8 | C | B | A | D |
| 9,10,11 | D | C | B | A |
| Par | A | B | C | D |

Interfaces to the gigabit transceivers will utilize the transceiver bus as a split bus with two separate routeword and data busses. The routeword bus will be a fixed size (2 bits for OC48 ingress, 4 bits for OC48 egress, 8 bits for OC192 ingress and 16 bits for OC192 egress), the data bus is a variable sized bus. The transmit order will always have routeword bits at fixed locations.

Every striping configuration has one transceiver that it used to talk to a destination in all valid configurations. That transceiver will be used to send both routeword busses and to start sending the data.

5 The backplane interface is physically implemented using 125 MHz interfaces to the backplane transceivers. The 125 MHz bus for both ingress and egress is viewed as being composed of two halves, each with routeword data. The two bus halves may have information on separate packets if the first bus half ends a packet.

10 For example, an OC48 interface going to the fabrics locally speaking has 24 data bits and 2 routeword bits @125 MHz. This bus will be utilized acting as if it has 2x (12 bit data bus + 1 bit routeword bus). The two bus halves are referred to as A and B. Bus A is the first data, followed by bus B. A packet can start on either bus A or B and end on either bus A or B.

15 In mapping data bits and routeword bits to transceiver bits, the bus bits are interleaved. This ensures that all transceivers should have the same valid/invalid status, even if the striping amount changes. Routewords should be interpreted with bus A appearing before bus B.

20 The bus A/Bus B concept closely corresponds to having 250 MHz interfaces between chips.

25 All backplane busses support fragmentation of data. The protocol used marks the last transfer (via the final segment bit in the routeword). All transfers which are not final segment need to utilize the entire bus width, even if that is not an even number of bytes. Any given packet must be striped to the same number of

THE UNIVERSITY OF CHICAGO

5

8 bit data bus, 1 bit clock, 1 bit control.

On the receive side, for channel the ASIC receives

a receive clock, 8 bit data bus, 3 bit status bus.

10.

20

25

Depending on the bus configuration, multiple channels may need to be concatenated together to form one larger bandwidth pipe (any time there is more than one transceiver in a logical connection. Although quad gbit transceivers can tie 4 channels together, this functionality is not used. Instead the receiving ASIC is responsible for synchronizing between the channels from one source. This is done in the same context as the generic synchronization algorithm.

The 8b/10b encoding/decoding in the gigabit transceivers allow a number of control events to be sent over the channel. The notation for these control events are K characters and they are numbered based on the encoded 10 bit value. Several of these K characters are used in the chipset. The K characters used and their functions are given in the table below.

TABLE 11: K Character usage

| K character | Function | Notes |
|-------------|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| 28.0 | Sync indication | Transmitted after lockdown cycles, treated as the prime synchronization event at the receivers |
| 28.1 | Lockdown | Transmitted during lockdown cycles on the backplane |
| 28.2 | Packet Abort | Transmitted to indicate the card is unable to finish the current packet. Current use is limited to a port card being pulled while transmitting traffic |
| 28.3 | Resync window | Transmitted by the striper at the start of a synch window if a resynch will be contained in the current sync window |
| 28.4 | BP set | Transmitted by the striper if the bus is currently idle and the value of the bp bit must be set. |
| 28.5 | Idle | Indicates idle condition |
| 28.6 | BP clr | Transmitted by the striper if the bus is currently idle and the bp bit must be cleared. |

The switch has a variable number of data bits supported to each backplane channel depending on the striping configuration for a packet. Within a set of transceivers, data is filled in the following order:



5 of this which is used for fill order is transceiver_number (OC48)
and transceiver number and oc48 port designation for OC192.

The fundamental rules for mapping are the following:

- # CCF-TL6

6. For OC192, the fill order should be bus A, B, C, D for routeword bits. For data bits, the fill order depends on wacking/unwacking/reverse unwacking and reverse wacking functions.

Transceiver 1

5 For an ingress bus, the format of data is the following:

Bit 0- BP
Bit 1- 0
Bit 2- RWA
Bit 3- RWB
Bit 4-Dataa(0)
Bit 5-Dataa(1)
Bit 6 Datab(0)
Bit 7 Datab(1)

Note that for 12 fabric mode, bits 5 and 7 are unused.

5 The location of datab(0) does not change.

For the egress bus, the format of the data is the following:

Bit 0- RWA(0)
Bit 1- RWA(1)
Bit 2- RWB(0)
Bit 3- RWB(1)
Bit 4-Dataa(0)
Bit 5-Dataa(1)
Bit 6 Datab(0)
Bit 7 Datab(1)

Transceiver 2 and up

Fill up the data bus starting at each transceiver bit 0 to bit 7 with 2 bit interleaved dataa/datab patterns.

5 For example, transceiver 2 has the following pattern:

Bit 0- dataa(2)

Bit 1- dataa(3)

Bit 2- datab(2)

Bit 3- datab(3)

10 Bit 4-Dataa(4)

Bit 5-Dataa(5)

Bit 6 Datab(4)

Bit 7 Datab(5)

The stop/align encoding depends on the width of the bus interface.

15 **TABLE 12:** OC48 portcard to fabric routeword stop/align

| Field | Length | Function |
|-------------------|-----------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Stop/Align /FS | 2 + n (where n is the number of clock cycles of transfer) | <p>In this mode, this field is stop & align & final_segment.</p> <p>Stop bit is a 1 to indicate no stop, zero indicates stop. Stop bits repeat in a serial stream until a stop bit of zero is seen, followed by the align bit and FS. Since stop is followed by the align and FS bits, the stop bit is given 2 clock cycles before the end of data.</p> <p>Align bit is a one to indicate valid data on the last complete byte on the interface. For odd 12 bit words (assuming zero based counting), align = 0 indicates bits 0:3 are valid, and bits 4:11 are invalid. Align = 1 for these words indicates that all 12 bits are valid. For even words, align should normally be a 1.</p> <p>Short packets are indicated by signaling a stop on byte 53 of the transfer. In reality, 54 bytes will be transferred, but the packet is flagged as a short packet.</p> <p>Final segment is a one to indicate a final segment of a packet and a zero to indicate a partial segment of a packet. Only one packet can be in transit at any one time on this bus. This bit is only valid for packets. For cells this bit should be a one. Packets which are not final segments should be terminated only on odd cycles with all bits utilized.</p> |
| | | |

TABLE 13: OC192 portcard to fabric routeword stop/align

| Field | Length | Function |
|-------------------|--------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Stop/Align /FS | 3 + 4 * number of extra clocks | <p>Due to length restrictions on this bus, the stop/align has to be treated differently than for OC48 transfers.</p> <p>The first clock cycle, this field is 3 bits long and is notated as SAF0. In all future clock cycles the stop field is 4 bits long and notated SAF1. The definitions of SAF0 and SAF1 are given below.</p> <p>SAF0(0). Bit zero is a zero to indicate a stop, a one to indicate no stop.</p> <p>SAF0(2:1)-"00" indicates full word transfer.</p> <p>"01" indicates a full word transfer but for a short packet.</p> <p>"10" indicates a full word transfer but not the final segment.</p> <p>"11" is reserved.</p> <p>SAF1(0) Bit zero is a zero to indicate a stop, a one to indicate no stop on the current cycle.</p> <p>SAF1(3:1)-binary value of the number of valid bytes. Zero is reserved and 7 is used to indicate 6 bytes valid but not the final segment. 6 indicates 6 bytes valid and final segment. All partial word transfers automatically indicate an implied final segment.</p> |

TABLE 14: OC48 Fabric-Port card routeword stop/align

| Field | Length | Function |
|-------------------|--------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Stop/Align /FS | 3 + 2 * number of extra clocks | <p>Value is treated as a repeated 2 bit value (encoded stop) followed by the final segment bit.</p> <p>Stop field is interpreted as:</p> <p>11-continue</p> <p>00-1st byte finished is valid and stop</p> <p>01-2nd bytes finished is valid and stop</p> <p>10-3rd byte finished is valid and stop, or non-final segment.</p> <p>Short packets are indicated by flagging a stop at byte 53.</p> <p>Final segment is a one for a final segment, a zero for a continuing packet. For final segments, the stop field should be encoded as a "10"</p> |

The port card - fabric interface at OC192 variable routeword bits are given in the table below.

TABLE 15: OC192 Fabric-port card routeword stop/align

| Field | Length | Function |
|------------|-------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Stop/Align | 7 + 8* number of extra clock cycles of transfer | <p>Bit 0 indicates stop. Zero indicates stop, 1 continue.</p> <p>Bits 4:1 give the number of valid bytes which complete on the interface if a stop is being executed.</p> <p>If no stop is being executed the value of these bits are don't cares. Zero is reserved. 0xC indicates 12 bytes and final segment. 0xD indicates the full bus and packet continues (not a final segment).</p> <p>Values 0xE, 0xF are reserved. Any non-12 byte ending offset automatically signals end of segment.</p> |

| | | |
|--|--|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| | | Bit 5:6 (first cycle) and bits 5:7 (second cycle and on) are reserved. The purpose of these bits is to align the next stop field with the following clock cycle of data. |
| | | Short packets are indicated by flagging a stop at byte 53. |
| | | |

Depending on the switch configuration, the bus may not transfer an integer number of bytes. This is handled by the interface always flagging the bytes which finish and the transmit and receive state machines must track where bytes begin and end based on the current cycle in the transfer.

The bus consists of a multiplexed address/data bus (AD_DATA), a select signal (AD_SEL_L), a read/write signal (AD_RW), and a bus transaction complete indication signal (AD_RDY_L). AD bus is used for read/write access of control/status registers.

In order to write to a control/status register, the read/write signal (AD_RW) must be low. The select signal (AD_SEL_L) must be asserted low for the entire duration of the access, and values must be placed on the AD_DATA bus in the following sequence (cycle 0 is the first cycle where AD_SEL_L is low for this transaction):

- cycle 2-5: Data to be written to control/status register. For registers that are wider than 8-bits (maximum of 32-bits) write data must be presented one byte per cycle starting with LSB. Any data presented on the bus beyond the width of the register will be ignored.
- cycles > 5: ASIC will assert AD_RDY_L on completion of the write access, and will keep it asserted until AD_SEL_L is de-asserted.

Figure 12 shows a Write Cycle.

In order to read from a control/status register, the read/write signal (AD_RW) must be high. The select signal (AD_SEL_L) must be asserted low for the entire duration of the access, and values must be placed on the AD_DATA bus in the following sequence (cycle 0 is the first cycle where AD_SEL_L is low for this transaction):

- cycle 0-1: Address of control/status register
- cycle 2: AD_DATA bus should be released (hi-z)
- cycles >3: When the data is available, ASIC will drive the read data onto the bus, one byte per cycle for four cycles, along with assertion of AD_RDY_L signal. For registers smaller than 32-bits wide, unused bits are presented as zeros. The LSB is present on the bus during the 1st clock cycle of AD_RDY_L assertion.

Figure 13 shows a Read Cycle.

The switch chips will generate interrupts on error conditions. The interrupt lines have the following characteristics:

1. Level Sensitive
2. Active Low

3. Asynchronous (no clock generated to go along with the interrupt).

4. Assume point-to-point interconnection with board logic which combines together interrupts.

5 Interrupts are maskable on a condition by condition basis inside each chip. The interrupt signal is asserted on the occurrence of an error condition and is cleared when the error condition is cleared. Any temporary conditions which caused an interrupt are recorded in the chip so no phantom interrupts should be seen.

The reality of the switch is that errors will occur. The intent in the following is to detail the expected system behavior and recovery strategy needed for each error type.

TABLE 16: Error recovery in the ASICs

| Error | Detection Mechanism | Error recovery required | Hardware comments |
|-------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Stuck bit on port card egress | unstriper sees data corruption from one fabric | | |
| Stuck bit between agg & memory controller | unstriper sees data corruption from one fabric, either route word or data. | | |
| Stuck bit between memory controller & separator | unstriper sees data corruption from one fabric, either route word or data | | |
| Stuck bit on fabric egress | | | |
| | | | |
| Soft-fail on routeword from port card | At least two unstripers see either a routeword mismatch, a state with a high number of routeword mismatches, or data parity errors or any number of unstripers will see a routeword mismatch, a high number of routeword mismatches or data parity errors and an aggregator will see a synch error. | Queue resynch | Worst case scenario involves failing routeword with different fabric routewords to fabrics. Either queueing a packet to the wrong port or dropping the traffic in the aggregator can cause an impact to all ports. Probability of impacting more ports goes up with traffic load and memory utilization in memory controllers. |

| | | | | |
|----|------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| | Soft-fail on data from port- card | Unstriper sees one time error, probability of automatic hardware based data recovery is high | None | |
| | Soft-fail between agg/memory controller dest_id bus | At least two unstripers see either a routeword mismatch, a state with a high number of routeword mismatches, or data parity errors | Queue resynch | |
| 5 | soft-fail between agg/memory controller data bus | Unstriper sees one time error, probability of automatic hardware based data recovery is high | None | |
| | soft-fail between memory controller/separator channel ID bus | At least two unstripers see either a routeword mismatch, a state with a high number of mismatches, or data parity errors | Queue resynch | Tokens get out of synch. May see error of FIFO overflow in the separator, depending on traffic pattern. Need congestion on the fabric for a port to have the FIFO overflow become possible. May also see excess tokens in memory controller. |
| 10 | soft-fail between memory controller/separator data bus for RW data | Packet boundaries from one separator port are lost. Unstriper will show a large number of errors for all traffic from the affected aggregator output. | Queue Resynch | Inherent that no self-stabilize in occurs w/o queue resynch. |
| | soft-fail between memory controller/separator data bus for packet data | Single port sees one-time error. | None | |
| 15 | soft-fail on token bus from separator to memory control ler | Mismatches from fabric due to differences in separator scheduling. | Queue Resynch | |
| | soft-fail internal to fabric chips | Unstriper sees different traffic from fabric than other fabrics | Reset | Queue Resynch may fix the problem, reset is necessary for restoring state. |
| | | | | |
| 20 | aggregator never sees back plane idle to synchronize to rw bus | Aggregator never sets flag indicating it has seen back plane sync | Replace faulty hardware. | same as below |
| | aggregator never sees system synch | Aggregator never sets flag indicating it has seen back plane sync | Replace Faulty hardware | Locating fault requires see in if only this board is hav in problems (backplane sync receiver) or if multiple boards are reporting problems (lost both sync signals on the back plane). Error isolation in 40G switch requires looking at the state of the secondary synch pulse generator |
| 25 | memory controller does not see synch from agg | | Retry resynch or if permanent replace faulty hardware. | |
| | separator does not see synch from mem_cont | Separator never gets initial synch | replace faulty hardware | |
| | unstriper does not see back synch | Unstriper never gets back plane synch | replace faulty hardware | |

| | | | |
|--------------------------------------------------------------------------|-----------------------------|------------------------|-----------------------------|
| | | | port card reset to recover. |
| soft failure on sync pulse distribution to multiple chips on a fabric | unknown | Reset the fabric | |
| soft failure on sync pulse distribution to multiple chips on a port card | Same as single-failure case | Same as single-failure | Same as single-failure. |
| | | | |

The chipset implements certain functions which are described here. Most of the functions mentioned here have support in multiple ASICs, so documenting them on an ASIC by ASIC basis does not give a clear understanding of the full scope of the functions required.

The switch chipset is architected to work with packets up to 64K + 6 bytes long. On the ingress side of the switch, there are busses which are shared between multiple ports. For most packets, they are transmitted without any break from the start of packet to end of packet. However, this approach can lead to large delay variations for delay sensitive traffic. To allow delay sensitive traffic and long traffic to coexist on the same switch fabric, the concept of long packets is introduced. Basically long packets allow chunks of data to be sent to the queueing location, built up at the queueing location on a source basis and then added into the queue all at once when the end of the long packet is transferred. The definition of a long packet is based on the number of bits on each fabric. The following table gives the size of long packets for different switch sizes.

TABLE 17: Long Packet sizes

| Switch Size | Packet Size (bytes) |
|-------------|------------------------|
| 40 | 900 |
| 80 | 1800 |

| | |
|-----|------|
| 120 | 2700 |
| 160 | 3600 |
| 240 | 5400 |
| 480 | 9600 |

5 If the switch is running in an environment where Ethernet MTU is maintained throughout the network, long packets will not be seen in a switch greater than 40G in size.

10 A wide cache-line shared memory technique is used to store cells/packets in the port/priority queues. The shared memory is 8K entries x 200-bit wide running at 125MHz. Each memory controller ASIC yields 25Gbps memory bandwidth. The aggregator #9 (control port) generates at most 4 streams of OC-48 traffic. The enqueue and dequeue speed for different switch configurations is shown in the following table. Note that a 2x speedup can be achieved for all switch configurations except the 480Gswitch. Up to 15 234,057 cells can be stored in the 480G switch. The shared memory stores cells/packets continuously so that there is virtually no fragmentation and bandwidth waste in the shared memory.

20 For the short packets/cells, memory utilization can be close to 100%. For the long packets, the memory block before the start of a long packet can be almost completely wasted. The minimum length for a long packet is 3 cache lines, giving an effective utilization of memory close to 75% since 1 out of 4 memory cache lines can be wasted.

25 **TABLE 18:** Shared Memory (1,638,400 bits) in Each Memory Controller

| Switches | En queue Speed | De queue Speed | Speed up Ratio | Cell Length | Number of Cells |
|----------|----------------|----------------|----------------|-------------|-----------------|
| 40G | 4.3Gbps | 20.7Gbps | 4.8 | 39+1 bits | 40,960 |
| 80G | 4.7Gbps | 20.3Gbps | 4.3 | 21+1 bits | 74,472 |
| 120G | 5.0Gbps | 20Gbps | 4 | 15+1 bits | 102,400 |

| | | | | | |
|------|---------|----------|-----|-----------|---------|
| 160G | 5.3Gbps | 19.7Gbps | 3.7 | 12+1 bits | 126,030 |
| 240G | 7Gbps | 18Gbps | 2.6 | 9+1 bits | 163,840 |
| 480G | 9.4Gbps | 15.6Gbps | 1.7 | 6+1 bits | 234,057 |

There exists up to 200 queues in the shared memory. They are per-destination and priority based. All cells/packets which have the same output priority and blade/channel ID are stored in the same queue. Cells are always dequeued from the head of the list and enqueued into the tail of the queue. Each cell/packet consists of a portion of the egress route word, a packet length, and variable-length packet data. Cell and packets are stored continuously, i.e., the memory controller itself does not recognize the boundaries of cells/packets for the unicast connections. The packet length is stored for MC packets. There is a limitation of 4K packets (or cells) in each of the MC queues.

The multicast port mask memory 64Kx16-bit is used to store the destination port mask for the multicast connections, one entry (or multiple entries) per multicast VC. The port masks of the head multicast connections indicated by the multicast DestID FIFOs are stored internally for the scheduling reference. The port mask memory is retrieved when the port mask of head connection is cleaned and a new head connection is provided.

Two configurations of port mask memory are supported:

- a. 8K port connections, for a 240 G switch
- b. 4K connections, for a 480 G switch.

Dequeue performance is restricted by several factors: 1) Padding injected by the aggregator ASICs; 2) Left alignment entries inserted in the memory controllers; 3) Memory controller output bus

fragmentation caused by the multicast connections; 4) Token bus latency between the separators and the memory controllers; 5) Separator output bus padding; and 6) Unstriper output bus fragmentation. A 480G switch is used as an example to analyze the worst-case performance since it has most padding, overhead, and congested traffic.

0 The aggregator ASICs have to pad a packet (including 36-bit route word, variable-length packet length field and datagram) to multiples of 12 since there are 12 memory controllers in one fabric. The shortest packet each memory controller received is 7-bit long since a packet can be as short as 84-bit long. The effective datagram is 3 bits. One entry will be left aligned for every 16 200-bit memory entries. The left aligned entry can be as short as 1-bit long. The worst-case datagram dequeue efficiency per output port of a memory controller is:

5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
110
115
120
125
130
135
140
145
150
155
160
165
170
175
180
185
190
195
200
205
210
215
220
225
230
235
240
245
250
255
260
265
270
275
280
285
290
295
300
305
310
315
320
325
330
335
340
345
350
355
360
365
370
375
380
385
390
395
400
405
410
415
420
425
430
435
440
445
450
455
460
465
470
475
480
485
490
495
500
505
510
515
520
525
530
535
540
545
550
555
560
565
570
575
580
585
590
595
600
605
610
615
620
625
630
635
640
645
650
655
660
665
670
675
680
685
690
695
700
705
710
715
720
725
730
735
740
745
750
755
760
765
770
775
780
785
790
795
800
805
810
815
820
825
830
835
840
845
850
855
860
865
870
875
880
885
890
895
900
905
910
915
920
925
930
935
940
945
950
955
960
965
970
975
980
985
990
995

$$(10\text{-bit (dout_me bus width)} * (3/7) \text{ (datagram length in a shortest packet)} * (15/16) \text{ (left-aligned overhead)}) * 250\text{MHz (output bus speed)} * 12 \text{ (number of memory controllers)} / 24 \text{ (number of output ports per separator)} = 502\text{Mbps}$$

20 The best-case output data bus bandwidth per separator channel is 2-bit * 250MHz, i.e., 500Mbps. In other words, The worst-case dequeue bandwidth of a memory controller is bigger than the best-case output bandwidth of a separator port. 2x speedup can be achieved through the twice wide output bus of the separators. One sync cycle will be fired on the output bus of the separator every 128 cycles.

The output bus of the unstriper ASIC is 64-bit wide at 100MHz. It can only carry one packet per cycle. In the worst-case, up to 56 bits are wasted per packet for an OC48 port.

APS stands for a Automatic Protection Switching, which is a SONET redundancy standard. To support APS feature in the switch, two output ports on two different port cards send roughly the same traffic. The memory controllers maintain one set of queues for an APS port and send duplicate data to both output ports.

To support data duplication in the memory controller ASIC, each one of 192 unicast queues has a programmable APS bit. If the APS bit is set to one, a packet is dequeued to both output ports. If the APS bit is set to zero for a port, the unicast queue operates at the normal mode. If a port is configured as an APS slave, then it will read from the queues of the APS master port. For OC48 ports, the APS port is always on the same OC48 port on the adjacent port card.

Port mirroring is similar to the APS except that any port can pair with any port. Only one pair of port mirroring ports are supported. A 16-bit port mirror register is used to identify the master and slave port involved in the port mirror operation. All ports are compared to the master portion (bit 15:8) of the register when dequeuing. Port mirror can be disabled. Note that a port can either have APS enabled or port mirroring enable, not both. The value of the port mirror register can be changed on-fly by the shadow registers.

The shared memory queues in the memory controllers among the fabrics might be out of sync (i.e., same queues among different memory controller ASICs have different depths) due to clock drifts or a newly inserted fabric. It is important to bring the fabric

queues to the valid and sync states from any arbitrary states. It is also desirable not to drop cells for any recovery mechanism.

5 A resync cell is broadcast to all fabrics (new and existing) to enter the resync state. Fabrics will attempt to drain all of the traffic received before the resynch cell before queue resynch ends, but no traffic received after the resynch cell is drained until queue resynch ends. A queue resynch ends when one of two events happens:

1. A timer expires.
2. The amount of new traffic (traffic received after the resynch cell) exceeds a threshold.

At the end of queue resynch, all memory controllers will flush any left-over old traffic (traffic received before the queue resynch cell). The freeing operation is fast enough to guarantee that all memory controllers can fill all of memory no matter when the resynch state was entered.

Queue resynch impacts all 3 fabric ASICs. The aggregators must ensure that the FIFOs drain identically after a queue resynch cell. The memory controllers implement the queueing and dropping. The separators need to handle memory controllers dropping traffic and resetting the length parsing state machines when this happens. For details on support of queue resynch in individual ASICs, refer to the chip ADSSs.

Multicast connections are enqueued into one of 4 priority queues based on the 2-bit priority number. They are stored cache-line based like the way unicast connections do. Connection numbers and lengths are stored into one of 4 1K-entry per-priority connection FIFO. Multicast packets are subject to be dropped if the

destined connection FIFO is full. In other words, at most 1K multicast packets can be stored simultaneously for each priority.

The 64Kx16-bit port mask memory will limit the number of multicast connections supported to 64K, 32K, 16K, 16K, 8K, and 4K for the 40G, 80G, 120G, 160G, 240G, and 480G switch, respectively.

For the dequeue side, multicast connections have independent 32 tokens per port, each worth up to 50-bit data or a complete packet. The head connection and its port mask of a higher priority queue is read out from the connection FIFO and the port mask memory every cycle (125MHz). A complete packet (or 50 bits if the packet is longer than 50 bits) is isolated from the 200-bit multicast cache line based on the length field of the head connection. The head packet is sent to all its destination ports. The 8 queue drainers transmit the packet to the separators when there are non-zero multicast tokens available for the ports. Next head connection will be processed only when the current head packet is sent out to all its ports.

For the worst case analysis, use the 480G switch as an example where the shortest packet is 7 bit long. Every 8ns cycle only one connection can be handled (bottlenecked by the connection FIFO and port mask memory). If the multicast only goes to 1 port, the effective dequeue throughput for the multicast connection is 875Mbps out of available 15Gbps shared memory dequeue bandwidth, i.e., 6%. In other words, the multicast performance is severely damaged by the bottlenecks existing in the connection FIFO, port mask memory, and head-of-line blocking. The throughput for the 480G switch is $480 \times 7 \times n / 80 = n \times 42G$ where n is number of copies a multicast connection destined. In the worst case where $n=1$, the multicast throughput is about 9% available switch capacity. If the average

multicast connections make 11 copies, the switch can achieve 480G throughput.

The longer a packet is (for the 240G switch or smaller configurations), the more ports a multicast connection destined, the dequeue performance becomes better significantly. Multicast performance do not intervene the dequeue speedup for unicast connections since the latter has their own tokens and two types of connections share the dout_me bus alternatively in a strict round-robin fashion, i.e., the multicast connections do not block unicast ones.

There are 192 unicast queues, 4 multicast queues, and 4 control port queues. 4 multicast queues are per priority based and can broadcast to any subset of 192 output ports and the 4 control ports.

There are up to 196 destination channels (192 blade channels and 4 control ports) for the 480G switch. Each destination has a one-to-one mapped unicast queue. 4 multicast queues can broadcast to any subsets of 192 regular ports indicated by the per-connection based port mask entry. An OC-192 port uses one out of 4 queue locations. Other three queues are unused. All 8-bit fabric queue ID field on the DestID bus is used to identify one of 196 ports. 2-bit priority field is unused.

For the 240G switch, Up to 100 destination channels exist (96 blade channels and 4 control ports). 96 unicast destination queues have 2 priority queues each. 4 multicast queues can broadcast to any subsets of 96 ports indicated by the per-connection based port mask entry. An OC-192 port uses one out of 4 queue locations. Other three queues are unused. Lower 7-bit queue ID is used to identify one of 100 ports and lower 1-bit of priority

field is used to identify one of two priority queues in each port. Other queue ID bit and priority bit is unused.

For the 160G switch, Up to 68 destination channels exist (64 blade channels and 4 control ports). 64 unicast destination queues have 2 priority queues each. There are 68 unused queues 4 multicast queues can broadcast to any subsets of 68 ports indicated by the per -connection based port mask entry. An OC-192 port uses one out of 4 queue locations. Other three queues are unused. Lower 7-bit queue ID is used to identify one of 100 ports and lower 1-bit of priority field is used to identify one of two priority queues in each port. Other queue ID bit and priority bit is unused.

For the 120G or smaller switch, Up to 52 destination channels exist (48 blade channels and 4 control ports). 48 unicast destination queues have 4 priority queues each. 4 multicast queues can broadcast to any subsets of 48 ports indicated by the per -connection based port mask entry. An OC-192 port uses one out of 4 queue locations. Other three queues are unused. Lower 6-bit queue ID is used to identify one of 52 ports and 2-bit priority field is used to identify one of 4 priority queues in each port. Other queue ID bits are unused.

Queue structure can be changed on fly through the fabric resync cell where the number of priority per port field is used to indicate how many priority queues each port has.

The stripper ASIC resides on the network blade. It has following features:

- Support packet/cell interfaces. Can accept up to 3 GB/sec of sustained traffic (3.2 GB/sec in bursts) of cells, frames, or a mix of cell and frame traffic.

- Generates fabric routeword for all fabrics in the switch
- Calculates data for the parity fabric and adds checksum to the end of each packet.
- Support switch configuration: 40G, 80G, 120G, 160G, 240G, and 480G
- 5 • Generates appropriate signals to interface directly to the transmit side of the Gbit transceivers.

10 The Striper takes BIB cell/packet format from the ingress port ASIC. For the ATM interface, the ASX cell format is accepted from the Vortex ASIC of the Poseidon chipset at 2.5Gbps for the channelized blade. It consists of 4-byte route word, 4-byte ATM cell header (without HEC byte), and 48-byte payload. 36-bit the switch route word can be generated based on the ASX route word provided by the Vortex ASIC.

15 The Striper ASIC consists of three major blocks: the switch route word generator, the switch payload & checksum generator, and the switch parity generator.

20 The switch payload generator forwards 4-byte ATM cell head, 48-byte ATM cell payload and 2-byte checksum to up to 12 switch fabrics and 1 spare fabric. The cell bus is 2x 12-bit wide running at 125MHz.

The Striper ASIC duplicates the packet/cell and transmits various fragments to the fabrics. 12 data output buses of the striper ASICs are connected to the data input buses of the aggregator ASICs on the fabrics as follows:

25 Figure 14 shows strip ASIC architecture.

TABLE 19: Data bus connectivity of the Striper ASIC of blade #1

| Data bus (DOUT_ST_1_ch_bu) | 40G (1 fabric) | 80G (2 fabrics) | 120G (3 fabrics) | 160G (4 fabrics) | 240G (6 fabrics) | 480G (12 fabrics) |
|----------------------------|----------------------------------|----------------------------------------|----------------------------------------|----------------------------------------|----------------------------------------|------------------------------------|
| Bus #1 | DIN_AG_1_1_ch_1 =cell[11:0] | DIN_AG_1_1_ch_1 [5:0]=cell[11:6] | DIN_AG_1_1_ch_1 [3:0]=cell[11:8] | DIN_AG_1_1_ch_1 [2:0]=cell[11:9] | DIN_AG_1_1_ch_1 [1:0]=cell[11:10] | DIN_AG_1_1_ch_1 [0]=cell[11] |
| Bus #2 | n/a | DIN_AG_2_1_ch_1 [5:0]=cell[5:0] | DIN_AG_2_1_ch_1 [3:0]=cell[7:4] | DIN_AG_2_1_ch_1 [2:0]=cell[8:6] | DIN_AG_2_1_ch_1 [1:0]=cell[9:8] | DIN_AG_2_1_ch_1 [0]=cell[10] |
| Bus #3 | n/a | n/a | DIN_AG_3_1_ch_1 =cell[3:0] | DIN_AG_3_1_ch_1 [2:0]=cell[5:3] | DIN_AG_3_1_ch_1 [1:0]=cell[7:6] | DIN_AG_3_1_ch_1 [0]=cell[9] |
| Bus #4 | n/a | n/a | n/a | DIN_AG_4_1_ch_1 [2:0]=cell[2:0] | DIN_AG_4_1_ch_1 [1:0]=cell[5:4] | DIN_AG_4_1_ch_1 [0]=cell[8] |
| Bus #5 | n/a | n/a | n/a | n/a | DIN_AG_5_1_ch_1 =cell[3:2] | DIN_AG_5_1_ch_1 [0]=cell[7] |
| Bus #6 | n/a | n/a | n/a | n/a | DIN_AG_6_1_ch_1 =cell[1:0] | DIN_AG_6_1_ch_1 [0]=cell[6] |
| Bus #7 | n/a | n/a | n/a | n/a | n/a | DIN_AG_7_1_ch_1 =cell[5] |
| Bus #8 | n/a | n/a | n/a | n/a | n/a | DIN_AG_8_1_ch_1 =cell[4] |
| Bus #9 | n/a | n/a | n/a | n/a | n/a | DIN_AG_9_1_ch_1 =cell[3] |
| Bus #10 | n/a | n/a | n/a | n/a | n/a | DIN_AG_10_1_ch_1 =cell[2] |
| Bus #11 | n/a | n/a | n/a | n/a | n/a | DIN_AG_11_1_ch_1 =cell[1] |
| Bus #12 | n/a | n/a | n/a | n/a | n/a | DIN_AG_12_1_ch_1 =cell[0] |
| Spare Fabric Bus | DIN_AG_sp_1_ch_1 parity[11:0] | DIN_AG_sp_1_ch_1 [5:0]= parity[5:0] | DIN_AG_sp_1_ch_1 [3:0]= parity[3:0] | DIN_AG_sp_1_ch_1 [2:0]= parity[2:0] | DIN_AG_sp_1_ch_1 [1:0]= parity[1:0] | DIN_AG_sp_1_ch_1 [0]= parity[0] |

The striper ASICs on blade #1 is connected with aggregator ASIC #1 of all switch fabrics. The striper ASICs on blade #2 is connected with aggregator ASIC #2 of all switch fabrics. The striper ASICs on blade #4 is connected with aggregator ASIC #4 of all switch fabrics. The striper ASICs on blade #5 to #8 are connected with aggregator ASIC #5 to #8 of all switch fabrics, respectively. The striper ASICs on blade #41 to #48 are connected with aggregator ASIC #5 to #8 of all switch fabrics, respectively. In other words, blade number moduled by 8 is the aggregator ASIC number which a striper ASIC is connected to.

The parity bits are sent to the spare fabric. The purpose of the spare fabric is to provide fault tolerance ability to the switch, i.e., in case one of the switch fabrics failed, the spare

fabric recovers the lost part of the cell. This is achieved through a parity bit generator on the striper ASIC. For one fabric configuration, the 12-bit cell payload is duplicated to the spare fabric; for 2-fabric configuration, 6-bit parity bits are generated as follows:

parity bit(1:6) = cell bit(1:6) exclusive-OR cell bit(7:12);

For 3-fabric configuration, 4-bit parity bits are generated as follows:

parity bit(1:4) = cell bit(1:4) exclusive-OR cell bit(5:8) exclusive-OR(9-12);

The route word generator regenerates the switch route word and sends up to 12+1 1-bit 250MHz route word buses for fabric 1,2,3,.., 12 and the spare fabric.

The aggregator ASIC resides on the switch fabric as shown in the following figure. Each 40G switch fabric has 8+1 aggregator ASICs. It aggregates 6x4 separate cell streams and route words into a single 12G stream from up to 6 blades and 4 channels. All input signals from the network blades are 250MHz point-to-point HSTL. It outputs a single cell stream that is multiplexed with cell payload and route words to 12 memory controllers. The ASIC has following features:

- 12Gbps Data and route word input from up to 6 network blades and 4 channels
- Route word separation and aggregation
- Output 12G data and route word to 12 memory controller ASICs

- HSTL interface with the memory controller, receiver interface for the backplane gigabit transceivers.

Figure 15 shows aggregator ASIC architecture.

The aggregator ASIC supports 40G, 80G, 120G, 160G, 240G, and 480G switch configuration without backplane change. The backplane connectivity (DIN_AG buses) of a pair of aggregator ASICs is shown as follows:

TABLE 20: DIN_AG bus connectivity of aggregator ASIC #1 and #5 of switch fabric #1

| DIN_AG_1_1_ch_bu | 40G (1 fabric) | 80G (2 fabrics) | 120G (3 fabrics) | 160G (4 fabrics) | 240G (6 fabrics) | 480G (12 fabrics) |
|----------------------|--------------------------------|--------------------------------|--------------------------------|---------------------------------|----------------------------------|-----------------------------|
| DIN_AG_1_5_ch_bu | | | | | | |
| DIN_AG_1_1_ch_1 | DOUT_ST_1_ch_1[5:0]=cell[11:0] | DOUT_ST_1_ch_1[5:0]=cell[11:6] | DOUT_ST_1_ch_1[3:0]=cell[11:8] | DOUT_ST_1_ch_1[2:0]=cell[11:9] | DOUT_ST_1_ch_1[1:0]=cell[11:10] | DOUT_ST_1_ch_1[0]=cell[11] |
| DIN_AG_1_5_ch_1[5:0] | n/a | DOUT_ST_5_ch_1[5:0]=cell[11:6] | DOUT_ST_5_ch_1[3:0]=cell[11:8] | DOUT_ST_5_ch_1[2:0]=cell[11:9] | DOUT_ST_5_ch_1[1:0]=cell[11:10] | DOUT_ST_5_ch_1[0]=cell[11] |
| DIN_AG_1_1_ch_2 | n/a | n/a | DOUT_ST_9_ch_1[3:0]=cell[11:8] | DOUT_ST_9_ch_1[2:0]=cell[11:9] | DOUT_ST_9_ch_1[1:0]=cell[11:10] | DOUT_ST_9_ch_1[0]=cell[11] |
| DIN_AG_1_5_ch_2[2:0] | n/a | n/a | n/a | DOUT_ST_13_ch_1[2:0]=cell[11:9] | DOUT_ST_13_ch_1[1:0]=cell[11:10] | DOUT_ST_13_ch_1[0]=cell[11] |
| DIN_AG_1_1_ch_3 | n/a | n/a | n/a | n/a | DOUT_ST_17_ch_1[1:0]=cell[11:10] | DOUT_ST_17_ch_1[0]=cell[11] |
| DIN_AG_1_5_ch_3 | n/a | n/a | n/a | n/a | DOUT_ST_21_ch_1[1:0]=cell[11:10] | DOUT_ST_21_ch_1[0]=cell[11] |
| DIN_AG_1_1_ch_4 | n/a | n/a | n/a | n/a | n/a | DOUT_ST_25_ch_1[0]=cell[11] |
| DIN_AG_1_5_ch_4 | n/a | n/a | n/a | n/a | n/a | DOUT_ST_29_ch_1[0]=cell[11] |
| DIN_AG_1_1_ch_5 | n/a | n/a | n/a | n/a | n/a | DOUT_ST_33_ch_1[0]=cell[11] |
| DIN_AG_1_5_ch_5 | n/a | n/a | n/a | n/a | n/a | DOUT_ST_37_ch_1[0]=cell[11] |
| DIN_AG_1_1_ch_6 | n/a | n/a | n/a | n/a | n/a | DOUT_ST_41_ch_1[0]=cell[11] |
| DIN_AG_1_5_ch_6 | n/a | n/a | n/a | n/a | n/a | DOUT_ST_45_ch_1[0]=cell[11] |

The 2 x 6 DIN_AG buses of aggregator ASIC #1 and #5 pair of switch fabric #1 is connected to the 12 x DOUT_ST bus #1 of blade #1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 41, and 45,

respectively. The 2 x 6 DIN_AG buses of aggregator ASIC #2 and #6 pair of switch fabric #1 is connected to the 12 x DOUT_ST bus #1 of blade #2, 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, and 46, respectively. The 2 x 6 DIN_AG buses of aggregator ASIC #3 and #7
5 pair of switch fabric #1 is connected to the 12 x DOUT_ST bus #1 of blade #3, 7, 11, 15, 19, 23, 27, 31, 35, 39, 43, and 47, respectively. The 2 x 6 DIN_AG buses of aggregator ASIC #4 and #8 pair of switch fabric #1 is connected to the 12 x DOUT_ST bus #1 of blade #4, 8, 12, 16, 20, 24, 28, 32, 36, 40, 44, and 48,
10 respectively.

Likewise, the 2 x 6 DIN_AG buses of aggregator ASIC #1 and #5 pair of switch fabric #2 is connected to the 12 x DOUT_ST bus #2 of blade #1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 41, and 45, respectively. The 2 x 6 DIN_AG buses of aggregator ASIC #1 and #5
15 pair of switch fabric #12 is connected to the 12 x DOUT_ST bus #12 of blade #1, 5, 9, 13, 17, 21, 25, 29, 33, 37, 41, and 45, respectively, for the 480G switch configuration.

The above connectivity is repeated 4 times for the channelized blades.

20 For the 40G, 80G, 120G, 160G, 240G, and 480G configuration, each blade channel sends 12 x 36-bit cell payload and 36-bit route word, 6 x 36-bit payload and 36-bit route word, 4 x 36-bit payload and 36-bit route word, 3 x 36-bit payload and 36-bit route word, 2 x 36-bit payload and 36-bit route word, and 1
25 x 36-bit payload and 36-bit route word to each switch fabric, respectively. In other words, the whole 12-bit wide cell is transmitted in the same fabric for the 40G switch while only a 1-bit wide (1/12 cell) cell slice is transmitted on each fabric for the 480G switch.

The 60-bit DOUT_AG bus is split onto 12 memory controller ASICs, each receiving 5-bit data and 1-bit clock signal from one aggregator ASIC. The 15-bit DestID bus is broadcast to all 12 memory controllers. Due to the fan out load concern, 3 copies of the signals are maintained, each driving 4 ASIC loads.

Every channel of the aggregator sends up to 12x3x200-bit cell/packet stream to 12 memory controller based on a work conserving round-robin dequeue algorithm, i.e., next source takes over if the current source runs out of eligible cells/packets to send. Strict round-robin algorithm is used among 24 sources. For the 40G switch, only 4 source channels exist. A source is eligible to send a cell/packet whenever a full cell or a full short packet or a 12x3x200-bit segment of a long packet is received.

Each memory controller ASIC receives 9 independent cell streams from 9 aggregator ASICs. There are 9 250MHz DIN_ME_fb_se buses, each consisting of a 5-bit data bus, a 1-bit clock signal, and a 15-bit DestID bus. The 60-bit DOUT_AG data buses of all 9 aggregator ASICs are bit sliced onto 12 memory controllers, each receiving 5-bit data from one DOUT_AG bus. Every memory controller gets a separate non-sharing clock signal (named clk1 to clk12) from each DOUT_AG bus to reduce the load of the clock pin while 3 memory controllers share a set of DestID bus from the DOUT_AG bus. The 9 DIN_ME_fb_se buses of memory controller #1 are connected to the DOUT_AG buses of 9 aggregators as follows:

- DIN_ME_fb_1_1_data = DOUT_AG_fb_1_data[48,36,24,12,0]
- DIN_ME_fb_1_1_dest = DOUT_AG_fb_1_dest1
- DIN_ME_fb_1_1_clk = DOUT_AG_fb_1_clk1
- DIN_ME_fb_1_2_data = DOUT_AG_fb_2_data[48,36,24,12,0]


```

5  °   DIN_ME_fb_1_2_dest = DOUT_AG_fb_2_dest1
    °   DIN_ME_fb_1_2_clk = DOUT_AG_fb_2_clk1
    °   DIN_ME_fb_1_3_data = DOUT_AG_fb_3_data[48,36,24,12,0]
    °   DIN_ME_fb_1_3_dest = DOUT_AG_fb_3_dest1
    °   DIN_ME_fb_1_3_clk = DOUT_AG_fb_3_clk1
    °   DIN_ME_fb_1_4_data = DOUT_AG_fb_4_data[48,36,24,12,0]
    °   DIN_ME_fb_1_4_dest = DOUT_AG_fb_4_dest1
    °   DIN_ME_fb_1_4_clk = DOUT_AG_fb_4_clk1
    °   DIN_ME_fb_1_5_data = DOUT_AG_fb_5_data[48,36,24,12,0]
10 °   DIN_ME_fb_1_5_dest = DOUT_AG_fb_5_dest1
    °   DIN_ME_fb_1_5_clk = DOUT_AG_fb_5_clk1
    °   DIN_ME_fb_1_6_data = DOUT_AG_fb_6_data[48,36,24,12,0]
    °   DIN_ME_fb_1_6_dest = DOUT_AG_fb_6_dest1
    °   DIN_ME_fb_1_6_clk = DOUT_AG_fb_6_clk1
15 °   DIN_ME_fb_1_7_data = DOUT_AG_fb_7_data[48,36,24,12,0]
    °   DIN_ME_fb_1_7_dest = DOUT_AG_fb_7_dest1
    °   DIN_ME_fb_1_7_clk = DOUT_AG_fb_7_clk1
    °   DIN_ME_fb_1_8_data = DOUT_AG_fb_8_data[48,36,24,12,0]
    °   DIN_ME_fb_1_8_dest = DOUT_AG_fb_8_dest1
20 °   DIN_ME_fb_1_8_clk = DOUT_AG_fb_8_clk1
    °   DIN_ME_fb_1_9_data = DOUT_AG_fb_9_data[48,36,24,12,0]
    °   DIN_ME_fb_1_9_dest = DOUT_AG_fb_9_dest1
    °   DIN_ME_fb_1_9_clk = DOUT_AG_fb_9_clk1

```

25 The DIN_ME data buses of memory controller #2 are
connected to bit 49,37,25,13, and 1 of the DOUT_AG data buses of 9

aggregators, and so on. The DIN_ME data buses of memory controller #12 are connected to bit 59,47,35,23, and 11 of the DOUT_AG data buses of 9 aggregators.

12 memory controller ASICs aggregate cell/packet streams from 8+1 aggregator ASICs. Then write the cells into one of 200 output queues (e.g., 12 network blades x 4 channelized Poseidon interfaces x 4 priorities for unicast + 4 priorities for multicast + 4 control port queues). The 8-bit destination queue number on the DestID bus is used as the output queue indicator for the unicast connection. The multicast cell is stored into one of 4 priority queues based on the 2-bit priority on the DestID bus. The 16-bit multicast connection number on the DestID bus will be used to lookup the internal port mask memory to find out the destination blade and channels during the dequeue phase.

The memory controllers send out cell/packet traffic from 200 output queues to 8+1 separator ASICs. Dequeueing speed is as twice fast as enqueueing speed to reduce amount of cells buffered on the switch fabric.

- Support both variable-length packet switching and fixed-length cell switching
- 12 ASICs are bit-sliced and function as an integrated shared memory controller
- Support 40G, 80G, 120G, 160G, 240G, and 480G switch configurations
- Enqueue cells/packets from 9 aggregator ASICs
- 2x dequeue speedup to 9 separator ASICs
- On-chip APS support
- 234,057 cells on-chip buffer

5 Figure 16 shows memory controller ASIC architecture.

Figure 17 shows wide cache line shared memory architecture.

DIN_ME_fb_se_9 and DOUT_ME_fb_se_9 buses are used to connect to aggregator #9 and separator #9, which communicate with the control port striper and unstriper ASICs only. It has the same DestID and cell format as other 8 buses do. Its cells are enqueued and dequeued in the same way as the regular cells.

There are up to 4 additional control port queues. They have queue ID from 192 to 195. All unicast connections having the control port queue ID as its fabric queue ID is enqueued into the relative control port queue. There are at most 4 OC-12 control ports supported.

Each control port queue has a 13-bit control port register as follows:

TABLE 21: 13-bit Control port queue register

| Bit 12:5 | Bit 4 | Bit 3 | Bit 2 | Bit 1 | Bit 0 |
|-----------------------|---------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 8-bit regular port ID | Regular Port enable | Control Port 3 enable | Control Port 2 enable | Control Port 1 enable | Control Port 0 enable |

A queue can be multicast to up to 4 physical control ports and one regular queue. When a queue is redirected to the regular queue, that queue must be disabled for the regular queue traffic. Packets are queued in the same way as the regular queues do, i.e., 200-bit cache line based. Left aligned every 16 cache lines. Strict round-robin among 4 queues when a left-alignment entry is transmitted. A queue is routed to 4 control ports and one regular port based on the 5-bit control port enable vector.

Two dequeue algorithms are applied among 4 control port queues:

- a) One control port only talks to one cp queue: Pure round-robin dequeue among 4 non-empty control port queues which have non-zero unicast tokens; one token worth unicast (up to 200-bit) is sent out to dout_me bus for a port;
- b) One control port talks to multicast cp queues: Strict priority among 4 control port queues; queue 192 has highest priority and queue 195 has lowest; switch queues when the end of the packet is seen.

OAM cells are identified by the Fabric queue ID field. If this field of a unicast connection has value 0xFx(h), then it is an OAM cell. All OAM cells can be mapped into one of the 192 blade or

4 control port queues set by a 8-bit programmable register (called OAM cell destination register).

Resync cell (0xFF) or any other special cells with fabric queue ID set to 0xFx are routed to any one of 196 queues based on the OAM cell destination register too.

Per destination minimum and maximum thresholds and counts can be set up to help memory management. 200x2x14-bit thresholds (in unit of 200-bit entry) and 200 x 13-bit running counters (in unit of 200-bit entry) are provided. Two additional per destination transmit and loss counts (32-bit each, in unit of packets) are also maintained. If the running count of a destination is above the relative threshold, new packets are rejected and loss count increments. Whenever dropping, the whole packet is dropped. Otherwise, the transmit count increments. For multicast connections, cells can also be rejected due to the multicast route word FIFO is full. 4 additional FIFO full counts are needed. If a packet is dropped, the whole packet is cleaned from the memory (including the segments of a long packet). The thresholds and current counts are in unit of 200-bit cache lines.

The minimum threshold (13-bit value plus 1-bit enable bit) is used to prevent shared memory starvation, i.e., every queue reserves at least the number of cache lines indicated by the threshold. The maximum threshold (13-bit value plus 1-bit enable bit) is used to prevent any single queue consuming the whole shared memory. These two thresholds cannot be changed unless there are no packets in the queues.

All counters are 32-bit wide. They are reset to zero automatically after reading. Their values stick to 0xFFFFFFFF if

- Route word separation and aggregation
- 0.25um 3V CMOS technology
- 410 I/O pins
- 140-bit 250MHz input; 240-bit 250MHz output (at most 120 of them switch simultaneously); 30-bit control signals

The separator has twice number of data output pins as that of the aggregator ASIC to support 2X speedup. Similar to those of the striper ASIC, the ASIC supports 40G, 80G, 120G, 160G, 240G, and 480G switch configurations without backplane change.

The separator ASIC performs reverse function of the aggregator ASIC. The ASIC receives 120-bit 250MHz cell/packet stream from one of 8 DOUT_ME_fb_se_bu buses of every memory controller (12 of them). 10-bit blade and channel selection signals are used to select one of 24 destinations inside each separator for up to two cells. For example, the DIN_SP buses of separator ASIC #1 is connected as follows:

- DIN_SP_fb_1_1=DOUT_ME_fb_1_1
- DIN_SP_fb_1_2=DOUT_ME_fb_2_1
- DIN_SP_fb_1_3=DOUT_ME_fb_3_1
- DIN_SP_fb_1_4=DOUT_ME_fb_4_1
- DIN_SP_fb_1_5=DOUT_ME_fb_5_1
- DIN_SP_fb_1_6=DOUT_ME_fb_6_1
- DIN_SP_fb_1_7=DOUT_ME_fb_7_1
- DIN_SP_fb_1_8=DOUT_ME_fb_8_1
- DIN_SP_fb_1_9=DOUT_ME_fb_9_1
- DIN_SP_fb_1_10=DOUT_ME_fb_10_1
- DIN_SP_fb_1_11=DOUT_ME_fb_11_1

- DIN_SP_fb_1_12=DOUT_ME_fb_12_1
- CH_SP_fb_1 = CH_ME_fb_1

When a valid cell/packet (channel ID is in the range of 0-23) is received, the packet type field in the route word is checked first. If it is an ATM cell, no packet length field is followed. The length of cell payload is 36x12/number of fabrics. If it is a packet, the packet length bit immediately followed is used to indicate how long a packet length is. 0=12-bit packet length (including this bit) and 1=24-bit packet length (including this bit). The entire packet/cell is routed to the destination channel indicated by the channel ID. The invalid channel ID (bigger than 24) is used to indicate that the cell/packet is invalid.

The ASIC then separate the route word and the payload onto the route word bus and the data bus of one of 6 blades and 4 destination channels/unstriper ASICs based on the channel ID signals. One 250MHz 24-bit data bus yields 6Gbps data bandwidth for each channel. Each route word is 2-bit wide running at 250MHz.

The connectivity between the separator ASICs and the Unstriper ASICs are symmetric to those between the aggregator ASICs and the striper ASICs. The only difference is that all data and route word pins have double-width to achieve 2X speedup.

Data received from each destination of each memory controller has a 1-bit valid bit accompanied. There are 24 destination input FIFOs are used to store the 12 pieces of cell/packets from 12 memory controllers for 24 destination blade and channels in each separator, respectively. When all 12 cell segments arrives, the complete cell is sent to the relative output FIFO indicated by the channel ID.

Like the striper ASIC, a 3-bit sequence number counter is maintained for the backplane synchronization. It increments every 36 250MHz cycles. When a cell is sent to the unstriper ASICs via the backplane, the current counter is attached into the sequence number field in the 36-bit route word.

The sequence number counter is reset by the global resynchronization logic.

The unstriper ASIC takes 6Gbps traffic from up to 12+1 switch fabrics. It then unstripes the cell and send it to the egress netmod ASIC at 5Gbps or lower speed.

- Receive 6Gbps route word and data from up to 12+1 fabrics at 250MHz for OC48 or combine 4 chips to support 20 Gbps routeword and data from up to 12+1 fabrics for OC192c
- Error check data transport throughout the switch, detect corrupted data and perform data recovery
- Reconstructs cells/packets from the individual switch fabrics.
- Send 64-bit 100MHz data to the egress port ASIC for OC48, 256 bit for OC192c
- Supports both UC and MC connection context for fabric data.

Figure 19 shows the unstriper ASIC Architecture.

The unstriper ASIC receives cells from up to 12+1 fabrics, each running at 250MHz. It uses the following steps to reconstruct good data.

1. All incoming routewords are compared. If any one routeword disagrees, that data lane is flagged as being in error. If more than one routeword disagrees, the data is dropped.

2. All valid input lanes are put through reconstruction logic which will attempt to build n+1 candidate output data streams for an N fabric switch. Any data lane which is not valid will invalidate any data lane which uses that data.

3. All valid reconstruction lanes will check the CRC of the received data and one passing output is selected.

The striper remaps the separate routeword and data buses to a combined outgoing routeword +data bus.

The following will detail the steps which happen at power up from an architectural perspective. Note that when expanding switch capacity, the additional fabrics must be brought on-line before any new port cards are brought on-line.

Fabric Initialization

1. Port cards (unstripers) are initialized to only look at current fabric capacity and ignore other fabric inputs.
2. Fabric is inserted, asserts its board present signal. Stripers start sending routewords to the new fabrics, though they are ignored at this point.
3. Board is reset, MCP starts to boot the board. Before proceeding to the next step, the MCP/SCP establish communication via the e-net network.
4. If the board is fabric 0 or the parity fabric, the sync pulse transmitter is initialized. (Actually sync pulse transmitter

can be initialized on all fabrics, but it is only connected to BP signals if it is fabric 0 or the parity fabric.)

5. MP initializes sync registers in the aggregator, memory controller, and separator, then initializes the registers in the sync pulse receiver. The sync pulse receiver starts to look for a valid sync pulse. The last sync setup is the sync pulse receiver, so that all receivers on the chips are ready for the sync pulse from the sync pulse receiver. The fabric chips run chip-chip sync on the next backplane sync pulse. The MP should check to make sure the fabric has synchronized. If sync has not been achieved, reset the fabric chips and re execute step 4.

6. SCP tells MP the current switch capacity window to use. This is actually going to correspond to the current switch capacity (does not count the capacity of the new fabric if switch capacity is being expanded).

7. MP initializes the backplane transceiver networks with the current switch capacity (both send and receive) and initializes all registers except the aggregator input enables. Any values used for configurable options (which ports are OC48/OC192, memory thresholds, etc) need to be communicated and initialized at this point. Certain registers are initialized based on the switch board slot, which needs to be known at this point. From a software perspective, the biggest register set which must be done is to update the port mask table in the memory controllers to match the port mask table from another switch fabric.

8. Aggregator input enables are set for the current switch capacity. This will start enqueueing traffic on this switch board. The aggregators will need to see a bus idle followed by an increment in the transmit sequence number before starting to actually receive data.

9. SCP sends a queue resync cell. On cell return, fabric queues

are now synchronized. However, no valid data is being enqueued in the new fabric(s) and the fabric outputs are being ignored.

10. All unstripers must be configured to start utilizing the new fabric. Since queues have been resynchronized, the fabric dequeuing should be synchronized and no errors should be seen. If errors are seen, clear them, return to step 8.

11. After all unstripers have been updated, SCP tells all port card MCPs to update stripe amount inside each of the striper ASICs. The change in striper configuration will start the switch utilizing the additional capacity.

12. After all stripe amounts are updated and traffic from the previous stripe amount drained from the switch, then the switch capacity needs to be updated. The only fixed time bound way of ensure traffic from the previous stripe amount is flushed is to execute a queue resync. If not all traffic has been flushed from the system with the previous stripe amount, the switch will drop this traffic at the unstripers (since there is no synchronization of the update at the separators, the drop cannot be performed there).

Before a port card is brought on-line, any necessary switch fabrics must be brought on-line first. As per the switch standard convention, port card installation happens in order.

1a. The starting state has sufficient switch capacity to support the new port card. Aggregators are currently configured to ignore the input from any new board.

1b. Port card is inserted and asserts its board present signal. Port card sees sync pattern received from the fabrics.

2. The sync pulse receiver is initialized. The port card starts looking for a valid sync pulse on the backplane.

4. Striper transmitter is set up for the appropriate number of destination fabrics and the Gbit network control is initialized. Before the GBit networks are initialized, the fabrics cannot count on seeing idle data from the new port card. At this point, the
5 port card can communicate its type (OC48/OC192) to the fabrics.

5a. Fabrics configure the port card type and enable the input from the port card.

5b. Striper/unstriper are now initialized, along with the other chips on the board. Some enable in the inbound data path should be disabled. The BIB input enable in the striper can be used or some other board specific input enable.

6. After both 5a and 5b have been completed, the port card can enable its input side and start sending data to the fabrics. Note that in general, further software configuration will need to be done after this point (such as setting up inbound lookup entries). The completion of 5a is necessary to ensure the fabric queues do not go out of sync.

7. First data from the port card is striped to all fabrics.

8. When a port card is removed from the system, not very much needs
20 to happen from a hardware perspective. Before the port card goes away, it transmits a packet abort which will cause any incomplete packets in the egress side to be dropped. Traffic will be drained from the memory queues which correspond to the affected output ports.

25 9. To remove a port card from the switch logically, software should disable the striper output bus.

Fabric deactivation is similar to fabric activation in reverse. The steps include:

1. Switch capacity is being removed. If port cards are present in the switch which are paired with the fabric capacity which is about
5 to be removed, those must first be deactivated.
2. Program the remaining stripers in the system to stripe data to one less stripe amount than the current configuration. This will stop sending real data to the fabric about to be decommissioned.
3. Send a queue resynch. This will flush out any traffic at the
10 last stripe amount.
4. Program the unstripers to start ignoring the data from the fabric which is about to be removed.
5. The fabric can now be physically removed from the system, or logically removed from the system by disabling its inputs and
15 outputs.

The reason for the queue resynch step is not because the switch is out of sync. The unstriper will treat the receipt of traffic which is striped to more fabrics than physically present in the switch as an error and increment error counts. The queue
20 resynch ensures that the error counts on the unstripers will not increment unnecessarily.

1. Flush out traffic from the port to be converted over to APS. Initialize anything in the separator as required for the new output port combination.

2. Write to the APS enable bit using the shadow register in every memory controller for the output port being affected. The main port for APS is not affected. Either a higher or lower number port can be the primary port and the backup port. APS is always enabled on
5 the backup port.

3. Send either a queue resync cell or a shadow control cell to all memory controllers.

4. Memory controllers start to dequeue after the next left-aligned cache boundary (if the previous transfer for this port was left-aligned, it will be remembered).
0

Note that in all this process, the queue number was never switched. The switch will not support a seamless port swap due to APS activate/deactivate. (In other words, APS can be turned on port 0, which will cause port 0 to mirror port 16. However, APS cannot be
5 turned off on port 16 since it is not on. Traffic is only being changed for the port where APS is added.)

The following words have reasonably specific meanings in the vocabulary of the switch. Many are mentioned elsewhere, but this is an attempt to bring them together in one place with definitions.
20

TABLE 22:

| Word | Meaning |
|-----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| APS | Automatic Protection Switching. A sonet/sdh standard for implementing redundancy on physical links. For the switch, APS is used to also recover from any detected port card failures. |
| Backplane | A generic term referring either to the general process the the switch boards use to account for varying transport delays between boards and clock drift or to the logic which implements the TX/RX functionality required for the the switch ASICs to account for varying transport delays and clock drifts. |
| 25 synch | |
| BIB | The switch input bus. The bus which is used to pass data to the striper(s). See also BOB |
| Blade | Another term used for a port card. References to blades should have been eliminated from this document, but some may persist. |
| BOB | The switch output bus. The output bus from the striper which connects to the egress memory controller. See also |

30 BIB.
Egress This is the routeword which is supplied to the chip after the unstriper. From an internal chipset perspective, the
Routeword egress routeword is treated as data. See also fabric routeword.
Fabric Routeword used by the fabric to determine the output queue. This routeword is not passed outside the unstriper.
Routeword A significant portion of this routeword is blown away in the fabrics.
Freeze Having logic maintain its values during lock-down cycles.

Lock-down Period of time where the fabric effectively stops performing any work to compensate for clock drift. If the
backplane synchronization logic determines that a fabric is 8 clock cycles fast, the fabric will lock down for 8
clocks.

35 Queue Resynch A queue resynch is a series of steps executed to ensure that the logical state of all fabric queues for all ports is
identical at one logical point in time. Queue resynch is not tied to backplane resynch (including lock-down) in
any fashion, except that a lock-down can occur during a queue resynch.
SIB Striped input bus. A largely obsolete term used to describe the output bus from the striper and input bus to the
aggregator.
SOB One of two meanings. The first is striped output bus, which is the output bus of the fabric and the input bus of
the agg. See also SIB. The second meaning is a generic term used to describe engineers who left Marconi to
form/work for a start-up after starting the switch design.
Sync Depends heavily on context. Related terms are queue resynch, lock-down, freeze, and backplane sync.
Wacking The implicit bit steering which occurs in the OC192 ingress stage since data is bit interleaved among stripers.
This bit steering is reversed by the aggregators.

40 The Aggregator Receive Synchronizer's function is to
maintain logical cell/packet ordering across all fabrics.
Cells/packets arriving at more than one fabric from different port
cards need to be processed in the same logical order across all
fabrics. If cell/packet logical ordering is not maintained, then
45 cells/packets coming out of fabrics will have stripes of a
particular cell/packet not match up and will not be able to be re-
assembled by the Unstriper.

Logical cell/packet ordering needs to be maintained
across the following conditions:

- 50 • Transport delay variances between one source and multiple
destinations
• Clock drift across transmitters and receivers

- Insertion and removal of port cards and fabrics
- Port card errors such as no sync, no lock-downs, too fast/too slow, routeword parity errors
- Gigabit transceiver errors such as loss-of-lock, data errors
- 5 • Non-synchronized updates to Gigabit network
- OC192c data streams (aggregating 4 channels to make up one OC192c stream)

0 The switch uses a system of transmit and receive
1 counters. The counters allow all components in the system to
2 logically align themselves. The Master Sequence Generator
3 implements these two counters that will count continuously from '0'
4 to '3' and will increment every x 125 MHz clock cycles where, x is
5 the counter tick length as programmed by software. x is currently
6 calculated to be 250 cycles. This is based on analysis done in the
7 Backplane Synchronization ADS. The relationship between the
8 transmit and receive counters can be seen in Figure 20. One
9 counter will be used by the transmit synchronizers in the Striper
10 and Separator ASICs and the other counter will be used in the
11 receive synchronizers in the Aggregator and Unstriper ASICs. The
12 receive counter will be a delayed version of the transmit counter.
13 The amount of delay is programmed by software in the Sync Pulse
14 Receive Delay register. This register determines the number of
15 clock cycles that the receive counter waits before incrementing its
16 own counter relative to the transmit counter. This register should
17 always be non-zero since the transmitter will have no delay and the
18 receiver needs to be delayed with respect to the transmitter. The
19 Sync Pulse Receive Delay has been estimated to be 150 cycles. The
20 delay is approximated equal to the worst case transport delay
21 between transmitter and receiver plus worst case transport delay
22 variance of the sync pulse. The delay also takes into account
23 worst case fast and slow transmitters and receivers.

The Sync Pulse Period is defined as the number of cycles between sync pulses. It is extended slightly by about 10 cycles in order for it to appear late in the `0' window of each ASIC's sequence count. This is done to ensure that every ASIC will appear to be running too fast even if they are actually running slow relative to the clock that generated the sync pulse. If this was not done, the sync pulse could appear in the `3' window and the ASIC would consider itself to be slow. There would be no way for it to catch up. Each transmitter and receiver will calculate the difference between when the sync pulse arrives and when its own counter transitions from `3' to `0'. This difference is the number of cycles that it is fast and is referred to as the lock-down amount (z in figure). Once a transmitter determines it should lock-down for z cycles, it will finish sending valid data during its `0' window and then lock-down z cycles. During the lock-down period, no valid or idle data is sent. Instead, a special lock-down K character is transmitted which will be recognized by the receiver. The receiver will not write the lock-down characters into its input FIFOs. This will ensure that the input FIFOs can't overflow. Since the sequence counter does not advance for the amount of lock-down, it is effectively resetting itself to the sync pulse. It is equivalent of having the sync pulse appear at the start of the `0' count window since the transition to a count of `1' occurs precisely one tick length after the sync pulse arrives. When the next sync pulse arrives, if clock frequencies are constant, then the sync pulse should appear in the `0' count window and the calculated lock-down amount will be the same as the previous calculation. This allows the system to always expect the sync pulse arrival in the `0' count window even if the clocks generating the sequence counter are too fast or too slow.

The Receive Synchronizer block will use the sequence counter to determine when to accept data from input byte sync FIFOs. Once a sync character is read, pops from the FIFOs will only occur once the sequence counter transitions from "0" to "1" and immediately following an arrival of a sync pulse. The read decision is only made once every sync pulse arrival and only at the "0" to "1" transition of the receive sequence counter. The sequence counter is also used during fabric resync in order to communicate a fabric resync to all channels in all aggregators during a sequence count transition. Fabric resync cells will be transmitted at the beginning of a sequence tick window and are prefixed by a special character indicating a resync cell. The receive synchronizers in the Aggregators will resynchronize all data going to the memory controllers on the next sequence count transition once the resync character has been received.

A block diagram of the receive Synchronizer can be seen in Figure 21. The Receive Synchronizer consists of 24 Byte-sync FIFOs, a Crossbar and 6 Bus Synchronizers. There is one byte sync FIFO per gigabit receiver. Each byte sync FIFO will accept data from each gigabit receiver independent of the mode of the switch. The byte sync FIFO depth is about 256 words deep. This depth is based on a derivation found in the Backplane Synchronizer ADS. The Crossbar will handle the assignment of the appropriate input byte lanes to the correct channels. Each Bus Synchronizer will consist of four Channel FIFOs and one Bus Controller. The Bus Controller can handle 4 separate OC48 channels or one OC192c stream. The channel FIFO is about 18 words deep. The depth is based on the number of words to read a 36-bit routeword. The whole routeword is read and then presented to the rest of the Aggregator in one cycle since it needs to be stored before the data of the packet as it is constructed and sent to the memory controller.

Multiple gigabit receivers make up a 24-bit data bus and 2-bit routeword bus for one channel of an Aggregator. Each gigabit receiver can handle up to 8 bits. Due to varying transport delays that can exist between receivers, bytes from different receivers that belong to the same word can be skewed from each other. For example, the 24-bit data bus and 2-bit routeword bus for one channel of an aggregator will have 4 receivers that make up the bus. The synchronization logic will align all 4 bytes for the 26-bit bus and will pass this byte aligned word to the rest of the Aggregator. In order to align the bytes, the Striper will need to send a special alignment byte to each receiver. A special K character can be utilized from the gigabit transceivers. The K character will be encoded in the data bits on the Gigabit transmitter and will be detected on the Gigabit receiver.

The receive synchronizer in the Aggregator will consist of 24 FIFOs where there is one FIFO per Gigabit Receiver. These FIFOs will handle both byte alignment and the backplane synchronization. It is assumed that the Gigabit Receivers will be able to distinguish between valid, idle, sync and lock-down cycles and will indicate these various cycles to the Aggregator by using 3 control signals.

On startup, the FIFOs will be empty and each Write State Machine (WSM) will wait until a sync character is seen on its input. From this point on, every cycle will be pushed except for lock-down cycles from the fabric. When the fabric is locking down, the Stripers will send special lock-down characters. This is done to avoid overflowing the sync FIFOs in case the write side clock is faster than the read side clock. While particular types of words are being pushed, the word type will also be written to the FIFO so it can be distinguished on the read side.

The WSM is also looking for a special fabric resync cell K character that will indicate that a fabric queue resync cell will immediately follow. If a resync cell is detected, a resync signal is passed along to Bus Controller. The Bus Controller will then
5 tell other Aggregators on the fabric to resync their queues at the next transition of the sequence counter. Fabric queue resync is described in more detail later.

Gigabit receivers are not dedicated to particular input channels, but instead shared between various channels. Each byte
10 sync FIFO works independently of the switch mode and each input lane needs to be steered to the correct channel FIFO. For instance in 40 mode, 26 bits of data and routeword are required for Bus 1, channel A and therefore 4 byte lanes are required to be steered to
15 each channel of Bus 1. In 80/120 mode, only 8 bits of data and 2 bits of routeword are required and therefore two bytes will suffice. In 480 mode, only 4 bits are required per channel and one byte lane will suffice. As switch capacity increases, less and less byte lanes will be required for a particular channel. For all
20 switch modes, the routeword bits for a particular channel will always come from the same byte lane. As the byte lanes get reduced from 4 to 1 byte lanes, there will always be one common byte lane used to carry the routeword data lines. The crossbar will take in
25 24 lanes consisting of 8 bits of data and 3 bits of control along with other control signals to communicate with the Bus Control logic. It will then forward all these signals to the appropriate channels. The Crossbar will also accept control data from the Bus Controller and forward signals such as read requests and FIFO flush signals to the appropriate input byte sync FIFOs. Each crossbar mapping between input byte lanes and channels is bi-directional.

30 The Bus Controller consists of three state machines. The state machines control the read side of the byte sync FIFOs, the

write side of the channel FIFOs and the read side of the Channel
FIFOs. On the read side of the Byte FIFOs, pops will not commence
until a sync pulse has arrived and the receive sequence counter has
transitioned from "0" to "1". A signal will be provided from the
5 sequence generator block that indicates a "0" to "1" transition at
precisely this moment(sync_event). At this time, the Bus
Controller issues a read to the Crossbar for the particular
channel. The Crossbar then forwards the read signal to the
appropriate byte sync FIFOs based on the mode of the switch. The
10 Crossbar then forwards all data and control from these byte sync
FIFOs back to the Bus Controller for this channel. The Bus
Controller checks the data types to make sure that the first word
in the appropriate byte sync FIFOs are a sync character. If the
first word of any of the appropriate byte lanes for this channel is
15 not a sync character, then a sync error will be flagged,
appropriate byte sync FIFOs will be flushed and the synchronization
process will be re-initiated. If the first word is a sync
character, then pops will continue. In OC48 mode, this process
will be performed independently for each channel. OC192c support
20 is discussed later on.

Once data starts being read from byte sync FIFOs, the Bus
Controller will ignore data until it finds the first idle word.
Once an idle word has been found, it can now start looking for the
SOP indication in the routeword when the next non-idle word is
25 read. The rest of the routeword is processed and made available to
the rest of the Aggregator. If the stop bit in the routeword
indicates that the packet is continuing, then data will be
continuously made available to the Aggregator until a stop
indication is read. Note that even though a SOP is seen, it does
30 not mean that this segment is the first segment of a packet. It
can be any segment of a packet. Even though the segment may not be

the first one of a packet, it is allowed to go through the switch and will be dropped later on.

When a sync character is read, a counter is initialized. The counter counts each read from the byte sync FIFOs. The Bus
5 Controller will expect to see a sync character every sync pulse period (about 22,000 cycles). If a sync character is read too early or too late, then a sync error is flagged, data is dropped at the precise logical cycle of where a sync character is expected. A packet that is being processed at the theoretical logical cycle for
10 sync will be terminated and inputs will be disabled until re-enabled by S/W. For example, if after the first sync character, the next sync character occurs at cycle 19,000, and then a sync error is flagged. Data is not dropped until 22,000 reads have been performed. Also, if after the first sync character, the next sync
15 character is not received at all after 22,000 cycles, then a sync error is flagged and data is dropped at this precise logical cycle. If a sync character is received precisely 22,000 cycles after the last one, then reads from the byte sync FIFOs are stopped until the receive sequence counter transitions from `0' to `1'. Waiting for
20 the `0' to `1' transition will ensure that all fabrics are receiving the same stripe of a packet on the same logical cycle.

For OC192c, 4 input channels need to be concatenated into one OC192c stream. In this mode, the Bus Controller will control all 4 channel FIFOs and the appropriate byte sync FIFOs. Data type
25 checking will be performed across 4 times as many byte lanes as in the OC48 case. When it is time to read byte sync FIFOs, the Bus Controller will control 4 read control lines to the Crossbar. The Crossbar will initiate reads across all appropriate byte sync FIFOs that are required for OC192c and will present data back to the Bus
30 Controller. The Bus Controller will check data types and will look for SOP indications. The SOP indication and stop bits will only be

000005067167000

Port card insertion

- 20 signal will be asserted and sent to each fabric. Not until S/W
enables the particular inputs and the Aggregator sees the port card
present signal, will the Aggregator be ready to accept data from
the new port card. Once enabled, the Aggregator will go through
the process of looking for sync characters on individual byte lanes
25 associated with the new port card. It is assumed that the port
card will not send any data until it has been configured only after
the fabrics have been initialized. Once the port cards are
enabled, they will start sending sync characters periodically at
every global sync pulse arrival. It is important that all the
30 appropriate fabrics see the sync character from the particular port

card since some fabrics will be initialized later than others. After sync characters have been received, all data will be written on each cycle excluding lock-down characters.

When a port card is about to be removed, the enable switch on the port card will be turned off. This will signal the port card to finish sending valid packets and then send idles. The port card will send a packet abort k character to indicate that no more valid packets will be sent immediately following the last valid packet. It is assumed that when the port card is actually removed, it will have already sent the packet abort k character. This is critical for the fabrics to keep their queues in sync. It is important that each Aggregator on each fabric that handles the particular port card stops forwarding data to the memory controllers at precisely the same logical cycle. The WSM will stop writing data into the byte sync FIFOs once the packet abort character is seen. The Bus Controller will terminate the packet once the packet abort character is read out of the byte sync FIFOs.

Case A: No sync/early sync/late sync from port card.

Solution: The Synchronizer will look for a sync at precisely the same logical cycle each time. This will occur every sync pulse period that is approximately 22,000 125MHz cycles. If the sync character is not present at the head of the byte sync FIFOs when 22,000 cycles have been read since the last sync character, a sync error will be flagged and data will be dropped the cycle where the sync character should have been. All fabrics need to drop data at precisely the same logical cycle for this particular input lane. Inputs for this particular channel will be turned off and the byte sync FIFOs used for this channel will be flushed. S/W will turn off the offending Striper. Inputs will be ignored until S/W enables these inputs again. If a sync character arrives too early, then data should be dropped at precisely the cycle where the early

sync was read. Other Aggregators will make the same drop decision if this error is common to all fabrics. If the sync character arrives too late or not at all, then the drop decision will be made where the sync character was expected. The sync character is
5 expected to arrive every 22,000 cycles after the last sync.

Case B: Port card not locking down.

Solution: If the port card does not lock-down, it will then send more than the ideal number of valid and idle cycles between sync characters. This will be caught by the same logic that checks for
10 sync characters in the correct logical cycles. Data will be dropped the same way as in the case where no sync came from the port card.

Case C: Routeword parity errors.

Solution: If a parity error is detected for a particular routeword,
15 the packet will be terminated at the bad segment and a parity error will be flagged. Data will be dropped after this terminated segment is forwarded to the rest of the Aggregator and FIFOs for this particular channel will be flushed. Inputs will be disabled until re-enabled by S/W.

20 Case D: Garbage data from port card while all fabrics already in sync.

Solution: If the data is unrecognizable by the gigabit receivers, errors will be formed and provided to the Aggregator by the gigabit receivers. At the point of error, data being written into byte
25 sync FIFOs will be flagged to be in error. If the Bus Controller sees that the particular byte lane in error is not used for the Routeword bits, then the error will be flagged but the data will be passed on to downstream logic. This is considered to be a soft failure since queues will still be able to stay in sync. If the
30 Bus Controller sees that the particular byte lane in error is used

for the Routeword bits, then the packet will be terminated and then dropped once the erred word is read from the byte sync FIFO. The input will be disabled, a gigabit receiver error will be flagged to S/W and byte sync and channel FIFOs associated with this channel will be flushed. This is considered to be a hard failure. If the failure occurs only for one fabric, then other fabrics can still be used to re-assemble the packets. S/W will have to queue resync the bad fabric. If this error occurs across multiple fabrics, not much can be done to avoid fabric queues from becoming corrupted. S/W will then have to queue resync all fabrics.

Case E: Port card sending data too fast or too slow. It is possible that the port card is sending the correct number of valid cycles between sync characters but is not locking down enough or locking down too much during each lock-down period. Byte sync FIFOs can eventually overflow or underflow respectively. If more than one fabric have FIFOs that overflow or underflow and data is dropped at different logical cycles for the same source, then fabric queues can become out of sync.

Solution: This is considered a hard failure since it should not occur if the hardware is working correctly. The only way to possibly prevent this is to flag an error if the FIFOs reach an almost full or almost empty threshold. This is a warning sign that something is wrong. S/W will then turn off the offending port card. Data will continue to be written to and read from the byte sync FIFOs as if nothing is wrong. If the port card can be turned off and idles be sent before byte sync FIFOs overflow, then there will be no dropped data and fabric queues will stay in sync. If FIFOs overflow or underflow for a particular channel, then a FIFO overflow/underflow error will be flagged. The packet being processed by the synchronizer at the time of error will be terminated. All data will be dropped from this point on. Inputs

for this channel will be disabled until re-enabled by S/W. FIFOs for this channel will be flushed.

5 Fabric queue resync is performed in order to resynchronize memory controller queues. It is important that all fabrics are processing the stripe of the same cell or packet at precisely the same logical cycle and that all fabrics are acting together as one logical fabric. Fabric queue resync starts at the Stripers. The Striper will receive a queue resync cell from the control port. The striper will decode the queue resync cell and will back up traffic until the next sequence counter tick is reached. At this point, it will send a fabric queue resync K character immediately followed by the queue resync cell. At the fabric, the WSM in the receive synchronizer will receive the queue resync K character and notify the Bus Controller in the receive synchronizer that a queue resync cell is in the input FIFO and that the queue resync event should occur at the next transition of the receive sequence counter. The Bus Controller will then indicate to other Aggregators on the fabric that a resync cell event will take place at the next transition of the sequence counter. The indication is asserted about 10 cycles before the receive sequence counter transitions. This is done to allow enough time for other Aggregators to see this assertion before their respective receive sequence counters transition also. Once the sequence count transition occurs, the Aggregators will signal to the memory controllers that a queue resync event has occurred and that this event delimits old and new data. All data sent before the sync event is considered old data and all data sent after the sync event is considered new data. The memory controllers synchronize their buffers accordingly. The resync cell is eventually sent through the switch as a regular cell and returned to the control port.

There can be times when the gigabit network is changing its operating mode and the switch is changing from a 40/80 to an 80/120 mode for example. There is no guarantee that Gigabit Receivers will be driven by Gigabit Transmitters during this time period. Aggregators that expect good data from certain Gigabit Receivers may not get good data. If the switch is increasing its mode, then a previously unused FIFO will now be used. If this FIFO has garbage data on its inputs, then syncs will not be received and this FIFO will not be synced until the gigabit network is stable. Once the Gigabit network is stable, idles and sync characters will be transmitted by the port cards and the FIFOs will have enough time to sync up. If the switch is decreasing its mode, then previously used FIFOs will now be unused. The Aggregator will know the new switch capacity and will eventually ignore these channel FIFOs.

The Unstriper needs to provide back-pressure to the Separators when internal FIFOs in the Unstriper become near full. Each Separator will expect 24 separate back-pressure signals coming from all the port card channels it is connected to. The back-pressure signal is considered to be asynchronous to all ASICs. It is required that all relevant Separators receive back-pressure from a particular channel in the Unstriper at precisely the same logical cycle. This is done by having the Unstripers assert the back-pressure signal when their receive sequence counter transitions. It is assumed that the Unstriper's receive sequence counter is a delayed version of the Stripers transmit sequence counter. Since the tick length is 250 cycles and the receive counter is delayed by 150 cycle relative to the transmit counter, there exists 100 cycles of margin to transport the back-pressure signal from the Unstriper to the Separator. The Separator needs about 10 cycles before the transition of its sequence counter to sample the back-pressure signal. This will give the Separator enough time to provide back-

pressure to the memory controller before the counter transitions. This places a maximum requirement on the propagation delay of the back-pressure signal. The following requirements hold true:

Back-pressure propagation delay < counter tick length - receive
5 sync pulse delay - setup time of Separator' sample point

Back-pressure propagation delay < 250 - 150 - 10

Back-pressure propagation delay < 90 cycles @ 125 MHz or 720 ns

Assuming worst-case conditions, the expected worst-case propagation delay would be:

0 Back-pressure propagation delay = (Unstriper to Striper delay) +
1 (Striper to Aggregator delay) + Aggregator to Separator Delay
2 Back-pressure propagation delay = 5 cycles (chip and board delay)
3 + (5+62 cycles) (chip and port card to fabric delay of 500 ns) + 5
4 cycles (chip and board delay)
5 Back-pressure propagation delay = 77 cycles < 90 cycles

As can be seen from this estimate, the maximum back-pressure propagation delay requirement is met.

Assuming all the relevant Separators receive the back-pressure signal before the transition to the next sequence
20 count, then it can be synchronized to the next transition of the
transmit sequence counter. This will allow all relevant Separators
to stop sending valid data at precisely the same logical cycle for
one complete counter tick interval. This is true since it is
assumed that when the transmit sequence counter transitions, the
25 data that the Separators are sending are companion fragments of the
same packet. If back-pressure is sampled again before the next
counter transition, then data will be stopped for another counter

tick interval. This mechanism implies that back-pressure can only be generated on a counter tick length granularity.

Since there is no direct path from Unstriper to Separator, the back-pressure signals need to be re-routed from the Unstriper, to the Striper, to the Aggregator and finally to the Separator. In order to do this, each Unstriper needs to send the back-pressure signal to the corresponding Striper on that port card. The Striper will then forward the back-pressure signal through the backplane gigabit transceivers onto the Aggregator. The Aggregator will forward up to 24 separate back-pressure signals to one Separator corresponding to 6 buses with 4 channels per bus. The back-pressure signal will always use bit 0 of the gigabit transceivers. The receive synchronizer block in the Aggregator will forward the correct back-pressure signal for the appropriate bus and channel to the Separator. Since the gigabit receivers are not dedicated to any particular bus and channel, the synchronizer needs to select the correct gigabit receiver based on the switch configuration just like it does for regular data. Once this is done, bit 0 of the gigabit receiver is forwarded on as the back-pressure signal. Note that bit 0 is also used for receiving k characters and can change when sending a k character. In order to avoid mistakenly interpreting bit 0 of a k character as a valid back-pressure signal, the synchronizer will only sample the back-pressure bit when valid data is received from the gigabit receiver. In the case where a k character is received, the synchronizer will hold the back-pressure signal at its current value. There is still a case where the Striper can be sending back-to-back idle characters since there is nothing to send. If the Striper needs to change the value of the back-pressure signal in this case, then it will send one of two k characters that change the back-pressure value. The two k characters that will be used are a set and clear of the back-pressure signal. If the synchronizer receives a

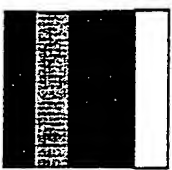
5

10

Figure 6

Notes:

1) Colors are used to represent the use/reuse of serial communication channels.



idle
committed
used in next config (off limits for this config)
2nd use of receiver/transmitter
3rd use of receiver/transmitter

2) Channel assignments are optimized for multiple variables:

- Minimization of I/O. Quad GE transceivers provide 1:2 fanout buffers and 2:1 multiplexers; 1:3 fanout or 3:1 fanin requires additional external devices.
- Partitioning of PCB routing complexity. The ASIC to quad GE buses are ideally straight and interleaved (striper with unstriper, and aggregator with separator) to simplify the routing of these clocked buses. The serial connections from quad GE transceivers to card edge connector on fabric and blade require significant untangling (with minimal use of vias) to group signals by common destination on the edge connector. This is required to reduce the complexity of the backplane routing and corresponding layer count.
- Maximization of symmetry between ASIC's. This symmetry allows interleaving of buses and reduces the need for external mux's and fanout buffers. It may also reduce the complexity of ASICs by reducing the number of unique mux/demux structures. Consequently, the striper map is a subset of the unstriper map, and both aggregator maps and the separator 5-8 map are derived from the separator 1-4 map.

Striper

| | | | | | | | |
|---------------|----|----|-----|-----|------|------|-----|
| # active tx | 8 | 6 | 8 | 10 | 7 | 7 | 13 |
| | 10 | | 10 | | 7 | | |
| | | 8 | | 12 | | 13 | |
| # fabrics | 1 | 2 | 3 | 4 | 6 | 6 | 12 |
| configuration | 40 | 80 | 120 | 160 | 240f | 240s | 480 |
| S_1A1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| S_1A2 | 2 | 2 | 2 | 2 | | | |
| S_1A3 | 11 | | | | | | |
| S_1A4 | 9 | | | | | | |

| | | | | | | | |
|--------|----|---|---|---|---|---|---|
| F1_1A1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| F1_1A2 | 4 | 4 | 4 | 4 | | | |
| F1_1A3 | 12 | | | | | | |
| F1_1A4 | 10 | | | | | | |

| | | | | | | |
|--------|---|---|---|---|---|---|
| F2_1A1 | 5 | 5 | 5 | 5 | 5 | 5 |
| F2_1A2 | 6 | 6 | 6 | | | |

| | | | | | |
|--------|---|---|---|---|---|
| F3_1A1 | 7 | 7 | 7 | 7 | 7 |
| F3_1A2 | 8 | 8 | | | |

| | | | | |
|--------|----|---|---|---|
| F4_1A1 | 9 | 9 | 9 | 9 |
| F4_1A2 | 10 | | | |

| | | | |
|--------|----|----|----|
| F5 1A) | JJ | 11 | 11 |
|--------|----|----|----|

| | | | |
|--------|----|----|----|
| F6_1A1 | 13 | 13 | 13 |
|--------|----|----|----|

F7_1A1 2

F8_1A1 4

F9_1A1 6

F10_1A1 8

F11_1A1 12

F12_1A1 14

| |
|----|
| 1 |
| 2 |
| 11 |
| 9 |

| |
|----|
| 3 |
| 4 |
| 12 |
| 10 |

| |
|---|
| 5 |
| 6 |

| |
|---|
| 7 |
| 8 |

| |
|----|
| 9 |
| 10 |

11

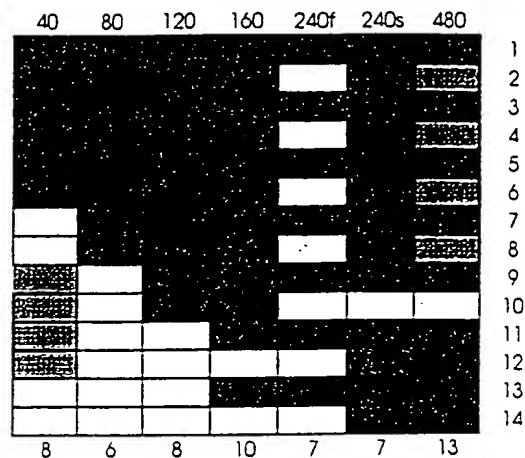
13

2

4

681214

Total (22)



8 internal 1:2 fanout buffers
6 internal 1:1 fanout buffers (one output can be disabled)

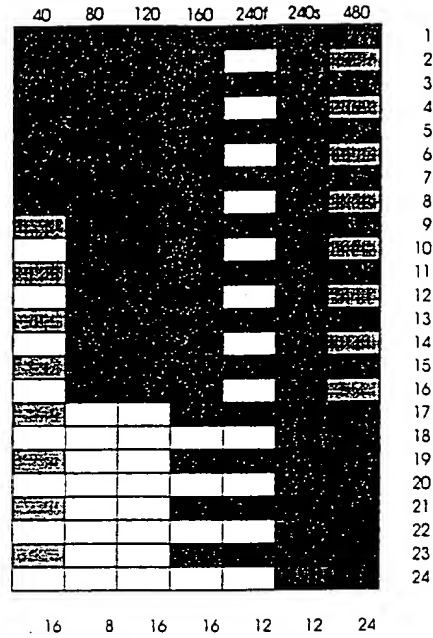
Aggregator 1-4

| # active rx | 16 | 8 | 16 | 16 | 12 | 12 | 24 |
|---------------|----|----|-----|-----|------|------|-----|
| | 16 | | 16 | | 12 | | 24 |
| # fabrics | 1 | 2 | 3 | 4 | 6 | 6 | 12 |
| Configuration | 40 | 80 | 120 | 160 | 240f | 240s | 480 |

| | | | | | | | | | |
|---------|-----|----|----|----|----|----|---|---|----|
| FI_1A1 | 1A | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| FI_1A2 | 1A | 2 | 2 | 2 | 2 | | | | 2 |
| FI_1A3 | 1A | 17 | | | | | | | 17 |
| FI_1A4 | 1A | 9 | | | | | | | 9 |
| FI_1B1 | 1B | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| FI_1B2 | 1B | 4 | 4 | 4 | 4 | | | | 4 |
| FI_1B3 | 1B | 19 | | | | | | | 19 |
| FI_1B4 | 1B | 11 | | | | | | | 11 |
| FI_1C1 | 1C | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| FI_1C2 | 1C | 6 | 6 | 6 | 6 | | | | 6 |
| FI_1C3 | 1C | 21 | | | | | | | 21 |
| FI_1C4 | 1C | 13 | | | | | | | 13 |
| FI_1D1 | 1D | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| FI_1D2 | 1D | 8 | 8 | 8 | 8 | | | | 8 |
| FI_1D3 | 1D | 23 | | | | | | | 23 |
| FI_1D4 | 1D | 15 | | | | | | | 15 |
| FI_9A1 | 9A | 9 | 9 | 9 | 9 | 9 | | | 9 |
| FI_9A2 | 9A | 10 | 10 | | | | | | 10 |
| FI_9B1 | 9B | 11 | 11 | 11 | 11 | 11 | | | 11 |
| FI_9B2 | 9B | 12 | 12 | | | | | | 12 |
| FI_9C1 | 9C | 13 | 13 | 13 | 13 | 13 | | | 13 |
| FI_9C2 | 9C | 14 | 14 | | | | | | 14 |
| FI_9D1 | 9D | 15 | 15 | 15 | 15 | 15 | | | 15 |
| FI_9D2 | 9D | 16 | 16 | | | | | | 16 |
| FI_17A1 | 17A | 17 | 17 | 17 | | | | | 17 |
| FI_17B1 | 17B | 19 | 19 | 19 | | | | | 19 |
| FI_17C1 | 17C | 21 | 21 | 21 | | | | | 21 |
| FI_17D1 | 17D | 23 | 23 | 23 | | | | | 23 |
| FI_25A1 | 25A | 2 | | | | | | | 2 |
| FI_25B1 | 25B | 4 | | | | | | | 4 |
| FI_25C1 | 25C | 6 | | | | | | | 6 |
| FI_25D1 | 25D | 8 | | | | | | | 8 |
| FI_33A1 | 33A | 10 | | | | | | | 10 |
| FI_33B1 | 33B | 12 | | | | | | | 12 |
| FI_33C1 | 33C | 14 | | | | | | | 14 |
| FI_33D1 | 33D | 16 | | | | | | | 16 |
| FI_41A1 | 41A | 18 | | | | | | | 18 |
| FI_41B1 | 41B | 20 | | | | | | | 20 |
| FI_41C1 | 41C | 22 | | | | | | | 22 |
| FI_41D1 | 41D | 24 | | | | | | | 24 |

increments by 8

40 total



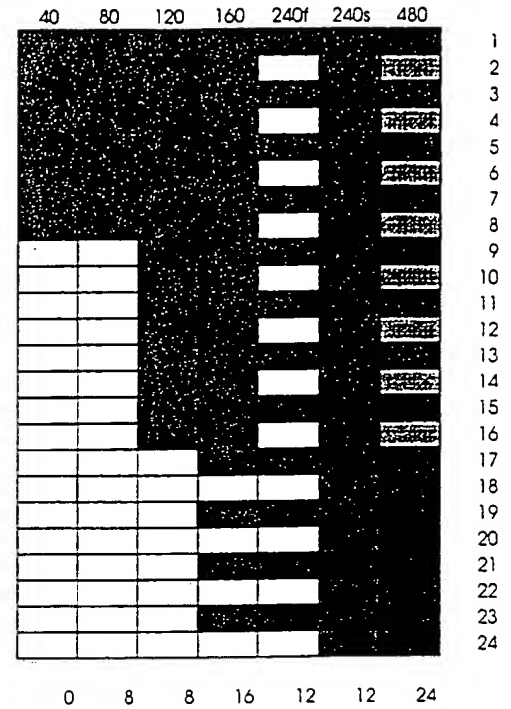
16 internal 2:1 Mux's

8 internal 1:1 buffers (one input can be turned off)

Aggregator 5-8

| # active rx | | 0 | 8 | 8 | 16 | 12 | 12 | 24 | |
|-------------|-----|----|----|-----|-----|------|------|-----|----|
| | | 8 | | 16 | | 12 | | 24 | |
| # fabrics | | 1 | 2 | 3 | 4 | 6 | 6 | 12 | |
| | | 40 | 80 | 120 | 160 | 240f | 240s | 480 | |
| F1_5A1 | 5A | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| F1_5A2 | 5A | 2 | 2 | 2 | | | | | 2 |
| F1_5B1 | 5B | 3 | 3 | 3 | 3 | 3 | 3 | | 3 |
| F1_5B2 | 5B | 4 | 4 | 4 | | | | | 4 |
| F1_5C1 | 5C | 5 | 5 | 5 | 5 | 5 | 5 | | 5 |
| F1_5C2 | 5C | 6 | 6 | 6 | | | | | 6 |
| F1_5D1 | 5D | 7 | 7 | 7 | 7 | 7 | 7 | | 7 |
| F1_5D2 | 5D | 8 | 8 | 8 | | | | | 8 |
| F1_13A1 | 13A | | | 9 | 9 | 9 | 9 | | 9 |
| F1_13A2 | 13A | | | 10 | | | | | 10 |
| F1_13B1 | 13B | | | 11 | 11 | 11 | 11 | | 11 |
| F1_13B2 | 13B | | | 12 | | | | | 12 |
| F1_13C1 | 13C | | | 13 | 13 | 13 | 13 | | 13 |
| F1_13C2 | 13C | | | 14 | | | | | 14 |
| F1_13D1 | 13D | | | 15 | 15 | 15 | 15 | | 15 |
| F1_13D2 | 13D | | | 16 | | | | | 16 |
| F1_21A1 | 21A | | | 17 | 17 | 17 | | | 17 |
| F1_21B1 | 21B | | | 19 | 19 | 19 | | | 19 |
| F1_21C1 | 21C | | | 21 | 21 | 21 | | | 21 |
| F1_21D1 | 21D | | | 23 | 23 | 23 | | | 23 |
| F1_29A1 | 29A | | | | | | 2 | | 2 |
| F1_29B1 | 29B | | | | | | 4 | | 4 |
| F1_29C1 | 29C | | | | | | 6 | | 6 |
| F1_29D1 | 29D | | | | | | 8 | | 8 |
| F1_37A1 | 37A | | | | | | 10 | | 10 |
| F1_37B1 | 37B | | | | | | 12 | | 12 |
| F1_37C1 | 37C | | | | | | 14 | | 14 |
| F1_37D1 | 37D | | | | | | 16 | | 16 |
| F1_45A1 | 45A | | | | | | 18 | | 18 |
| F1_45B1 | 45B | | | | | | 20 | | 20 |
| F1_45C1 | 45C | | | | | | 22 | | 22 |
| F1_45D1 | 45D | | | | | | 24 | | 24 |

32 total



8 internal 2:1 Mux's

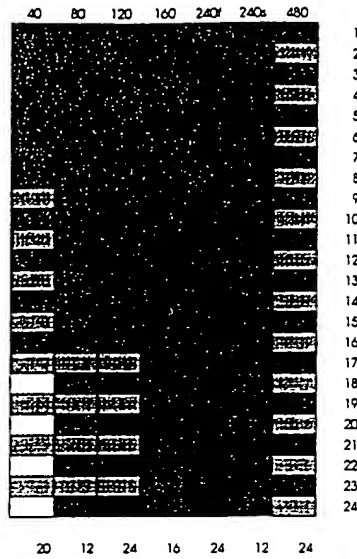
16 internal 1:1 buffers (one input can be turned off)

| Separator 1-4 | | | | | | | |
|---------------|----|----|-----|-----|------|------|-----|
| # active tx | 20 | 12 | 24 | 16 | 24 | 12 | 24 |
| # boards | 1 | 2 | 3 | 4 | 6 | 6 | 12 |
| Configuration | 40 | 80 | 120 | 160 | 240f | 240s | 480 |

| | | | | | | | | |
|---------|-----|----|----|----|----|----|---|---|
| FI_1A1 | 1A | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| FI_1A2 | 1A | 2 | 2 | 2 | 2 | 2 | | |
| FI_1A3 | 1A | 17 | 17 | 17 | | | | |
| FI_1A4 | 1A | 9 | | | | | | |
| FI_1A5 | 1A | 10 | | | | | | |
| FI_1B1 | 1B | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| FI_1B2 | 1B | 4 | 4 | 4 | 4 | 4 | | |
| FI_1B3 | 1B | 19 | 19 | 19 | | | | |
| FI_1B4 | 1B | 11 | | | | | | |
| FI_1B5 | 1B | 12 | | | | | | |
| FI_1C1 | 1C | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| FI_1C2 | 1C | 6 | 6 | 6 | 6 | 6 | | |
| FI_1C3 | 1C | 21 | 21 | 21 | | | | |
| FI_1C4 | 1C | 13 | | | | | | |
| FI_1C5 | 1C | 14 | | | | | | |
| FI_1D1 | 1D | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| FI_1D2 | 1D | 8 | 8 | 8 | 8 | 8 | | |
| FI_1D3 | 1D | 23 | 23 | 23 | | | | |
| FI_1D4 | 1D | 15 | | | | | | |
| FI_1D5 | 1D | 16 | | | | | | |
| FI_9A1 | 9A | 9 | 9 | 9 | 9 | 9 | | |
| FI_9A2 | 9A | 10 | 10 | 10 | | | | |
| FI_9A3 | 9A | 18 | | | | | | |
| FI_9B1 | 9B | 11 | 11 | 11 | 11 | 11 | | |
| FI_9B2 | 9B | 12 | 12 | 12 | | | | |
| FI_9B3 | 9B | 20 | | | | | | |
| FI_9C1 | 9C | 13 | 13 | 13 | 13 | 13 | | |
| FI_9C2 | 9C | 14 | 14 | 14 | | | | |
| FI_9C3 | 9C | 22 | | | | | | |
| FI_9D1 | 9D | 15 | 15 | 15 | 15 | 15 | | |
| FI_9D2 | 9D | 16 | 16 | 16 | | | | |
| FI_9D3 | 9D | 24 | | | | | | |
| FI_17A1 | 17A | 17 | 17 | 17 | | | | |
| FI_17A2 | 17A | 18 | | | | | | |
| FI_17B1 | 17B | 19 | 19 | 19 | | | | |
| FI_17B2 | 17B | 20 | | | | | | |
| FI_17C1 | 17C | 21 | 21 | 21 | | | | |
| FI_17C2 | 17C | 22 | | | | | | |
| FI_17D1 | 17D | 23 | 23 | 23 | | | | |
| FI_17D2 | 17D | 24 | | | | | | |
| FI_25A1 | 25A | | | | 2 | | | |
| FI_25B1 | 25B | | | | 4 | | | |
| FI_25C1 | 25C | | | | 6 | | | |
| FI_25D1 | 25D | | | | 8 | | | |
| FI_33A1 | 33A | | | | 10 | | | |
| FI_33B1 | 33B | | | | 12 | | | |
| FI_33C1 | 33C | | | | 14 | | | |
| FI_33D1 | 33D | | | | 16 | | | |
| FI_41A1 | 41A | | | | 18 | | | |
| FI_41B1 | 41B | | | | 20 | | | |
| FI_41C1 | 41C | | | | 22 | | | |
| FI_41D1 | 41D | | | | 24 | | | |

increments by 8

52 total



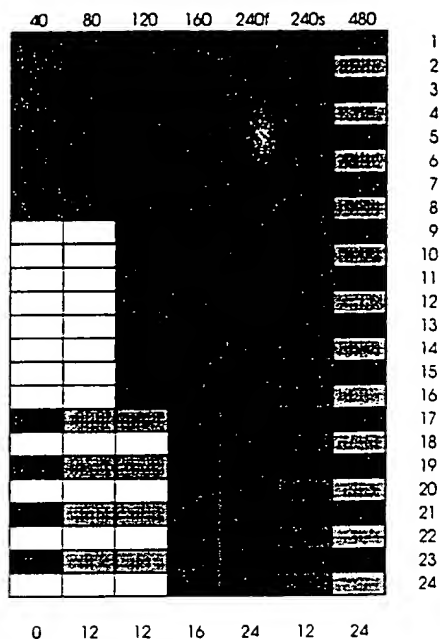
8 external 1:2 fanout buffers
20 internal 1:2 fanout buffers
4 internal 1:1 buffers (one output can be turned off)

Separator 5-8

| # active tx | 0 | 12 | 12 | 16 | 24 | 12 | 24 |
|--------------|----|----|-----|-----|------|------|-----|
| | 12 | | 20 | | 24 | | 24 |
| | | 12 | | 24 | | 24 | |
| # fabrics | 1 | 2 | 3 | 4 | 6 | 6 | 12 |
| Configuratia | 40 | 80 | 120 | 160 | 240f | 240s | 480 |

| | | | | | | | | |
|---------|-----|----|----|----|----|----|---|----|
| F1_5A1 | 5A | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| F1_5A2 | 5A | 2 | 2 | 2 | 2 | | | 2 |
| F1_5A3 | 5A | 17 | 17 | | | | | 17 |
| F1_5B1 | 5B | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| F1_5B2 | 5B | 4 | 4 | 4 | 4 | | | 4 |
| F1_5B3 | 5B | 19 | 19 | | | | | 19 |
| F1_5C1 | 5C | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| F1_5C2 | 5C | 6 | 6 | 6 | 6 | | | 6 |
| F1_5C3 | 5C | 21 | 21 | | | | | 21 |
| F1_5D1 | 5D | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| F1_5D2 | 5D | 8 | 8 | 8 | 8 | | | 8 |
| F1_5D3 | 5D | 23 | 23 | | | | | 23 |
| F1_13A1 | 13A | | 9 | 9 | 9 | 9 | | 9 |
| F1_13A2 | 13A | | 10 | 10 | | | | 10 |
| F1_13B1 | 13B | | 11 | 11 | 11 | 11 | | 11 |
| F1_13B2 | 13B | | 12 | 12 | | | | 12 |
| F1_13C1 | 13C | | 13 | 13 | 13 | 13 | | 13 |
| F1_13C2 | 13C | | 14 | 14 | | | | 14 |
| F1_13D1 | 13D | | 15 | 15 | 15 | 15 | | 15 |
| F1_13D2 | 13D | | 16 | 16 | | | | 16 |
| F1_21A1 | 21A | | 17 | 17 | 17 | | | 17 |
| F1_21A2 | 21A | | 18 | | | | | 18 |
| F1_21B1 | 21B | | 19 | 19 | 19 | | | 19 |
| F1_21B2 | 21B | | 20 | | | | | 20 |
| F1_21C1 | 21C | | 21 | 21 | 21 | | | 21 |
| F1_21C2 | 21C | | 22 | | | | | 22 |
| F1_21D1 | 21D | | 23 | 23 | 23 | | | 23 |
| F1_21D2 | 21D | | 24 | | | | | 24 |
| F1_29A1 | 29A | | | | | 2 | | 2 |
| F1_29B1 | 29B | | | | | 4 | | 4 |
| F1_29C1 | 29C | | | | | 6 | | 6 |
| F1_29D1 | 29D | | | | | 8 | | 8 |
| F1_37A1 | 37A | | | | | 10 | | 10 |
| F1_37B1 | 37B | | | | | 12 | | 12 |
| F1_37C1 | 37C | | | | | 14 | | 14 |
| F1_37D1 | 37D | | | | | 16 | | 16 |
| F1_45A1 | 45A | | | | | 18 | | 18 |
| F1_45B1 | 45B | | | | | 20 | | 20 |
| F1_45C1 | 45C | | | | | 22 | | 22 |
| F1_45D1 | 45D | | | | | 24 | | 24 |

40 total



16 internal 1:2 fanout buffers
12 internal 1:1 buffers (one output can be turned off)

Unstriper

| # active rx | 10 | 9 | 12 | 10 | 14 | 7 | 13 |
|---------------|----|----|-----|-----|------|------|-----|
| | 13 | | | 14 | | 14 | |
| | | 12 | | 14 | | | 13 |
| # fabrics | 1 | 2 | 3 | 4 | 6 | 6 | 12 |
| Configuration | 40 | 80 | 120 | 160 | 240f | 240s | 480 |

| | | | | | | | |
|-------|----|----|----|---|---|---|---|
| S_1A1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| S_1A2 | 2 | 2 | 2 | 2 | 2 | | |
| S_1A3 | 11 | 11 | 11 | | | | |
| S_1A4 | 9 | | | | | | |
| S_1A5 | 7 | | | | | | |

| | | | | | | | |
|--------|----|----|----|---|---|---|---|
| F1_1A1 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| F1_1A2 | 4 | 4 | 4 | 4 | 4 | | |
| F1_1A3 | 12 | 12 | 12 | | | | |
| F1_1A4 | 10 | | | | | | |
| F1_1A5 | 8 | | | | | | |

| | | | | | | | |
|--------|----|----|---|---|---|---|--|
| F2_1A1 | 5 | 5 | 5 | 5 | 5 | 5 | |
| F2_1A2 | 6 | 6 | 6 | 6 | | | |
| F2_1A3 | 13 | 13 | | | | | |

| | | | | | | | |
|--------|----|---|---|---|---|--|--|
| F3_1A1 | 7 | 7 | 7 | 7 | 7 | | |
| F3_1A2 | 8 | 8 | 8 | | | | |
| F3_1A3 | 14 | | | | | | |

| | | | | | | | |
|--------|----|----|---|---|--|--|--|
| F4_1A1 | 9 | 9 | 9 | 9 | | | |
| F4_1A2 | 10 | 10 | | | | | |

| | | | | | | | |
|--------|----|----|----|--|--|--|--|
| F5_1A1 | 11 | 11 | 11 | | | | |
| F5_1A2 | 12 | | | | | | |

| | | | | | | | |
|--------|----|----|----|--|--|--|--|
| F6_1A1 | 13 | 13 | 13 | | | | |
| F6_1A2 | 14 | | | | | | |

| | | | | | | | |
|--------|--|--|--|---|--|--|--|
| F7_1A1 | | | | 2 | | | |
|--------|--|--|--|---|--|--|--|

| | | | | | | | |
|--------|--|--|--|---|--|--|--|
| F8_1A1 | | | | 4 | | | |
|--------|--|--|--|---|--|--|--|

| | | | | | | | |
|--------|--|--|--|---|--|--|--|
| F9_1A1 | | | | 6 | | | |
|--------|--|--|--|---|--|--|--|

| | | | | | | | |
|---------|--|--|--|---|--|--|--|
| F10_1A1 | | | | 8 | | | |
|---------|--|--|--|---|--|--|--|

| | | | | | | | |
|---------|--|--|--|----|--|--|--|
| F11_1A1 | | | | 12 | | | |
|---------|--|--|--|----|--|--|--|

| | | | | | | | |
|---------|--|--|--|----|--|--|--|
| F12_1A1 | | | | 14 | | | |
|---------|--|--|--|----|--|--|--|

| |
|----|
| 1 |
| 2 |
| 11 |
| 9 |
| 7 |

| |
|----|
| 3 |
| 4 |
| 12 |
| 10 |
| 8 |

| |
|----|
| 5 |
| 6 |
| 13 |

| |
|----|
| 7 |
| 8 |
| 14 |

| |
|----|
| 9 |
| 10 |

| |
|----|
| 11 |
| 12 |

| |
|----|
| 13 |
| 14 |

| |
|---|
| 2 |
|---|

| |
|---|
| 4 |
|---|

| |
|---|
| 6 |
|---|

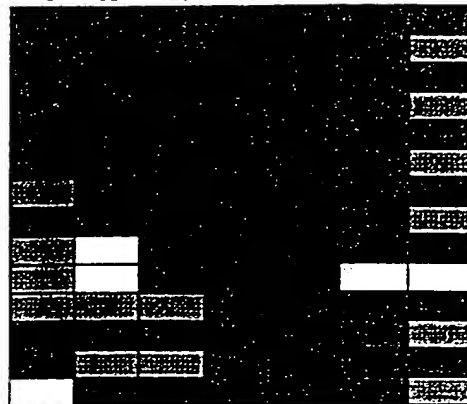
| |
|---|
| 8 |
|---|

| |
|----|
| 12 |
|----|

| |
|----|
| 14 |
|----|

Total 28

40 80 120 160 240f 240s 480



3 external 2:1 Mux's

11 internal 2:1 Mux's

3 internal 1:1 input buffers (one input can be disabled)

09610115-063000

BFS ASIC Backplane Connection Map

| Revision | Date | Comments |
|----------|----------|--------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | 11/3/99 | First release, checked by Spelser |
| 2 | 11/8/99 | Changed 40G assignments for channels 7 and 9; affects the striper and unstriper only. |
| | | Changed 40, 80, 120G assignments for channels 17-24 to make Separator 1-4 and Separator 5-8 symmetric (Separator 5-8 is now a subset of Separator 1-4) |
| 3 | 11/30/99 | Checked by Spelser and Blanchini |
| | | Added control port assignments. |

| | | | | | |
|------------------------------------------------------------------------------|------------|---------|------|----------|--|
| Fabric to blade egress bus format: <fabric>_<blade><channel><lane><polarity> | | | | | |
| Fabric slot | Blade slot | Channel | Lane | Polarity | |
| F1-F12,S | 1-48 | A-D | 1-5 | P,N | |

Naming convention:

Notes:

- 1) Map shown for a single instance of each ASIC, e.g. striper channel A for blade number 1. From an ASIC point of view, all that is significant in this striper table is the fabric number and the lane number; the topology is the same for each subsequent channel (B-D) blade slot (2-48).
- 2) Polarity is required to distinguish the two physical wires forming a differential pair. For each table entry (i.e., each logical signal) there are two physical connections. For example, logical signal F2_1A3 is really comprised of F2_1A3P and F2_1A3N, which connect fabric 2 to blade 1, unstriper 1, lane 3. Polarity is not shown in the connection maps to make them easier to read.
- 3) How to use this map: For a given configuration (e.g. with one fabric installed, read down 40 G column) Read down the column to determine the source for data present on an striper output bus. Each row represents a single striper output bus.
- 4) XXXXXXXXXX = not used in this Configuration
- 5) To aid in PCB routing and reduce design complexity, there are only two unique connection assignments, the unstriper and the Separator 1-4 maps. The striper map is a subset of the unstriper map. The separator 5-8 map, the aggregator 1-4 map, and the aggregator 5-8 map are all subsets of the separator 1-4 map.
- 6) Up to four control ports are supported (see maps for aggregator 9 and separator 9). If not all four control ports are needed, some control port channels C and D can be disabled.

Stripper

| Configuration | | | | | | | Stripper Output bus |
|---------------|--------|--------|--------|--------|--------|---------|---------------------------|
| 40 | 80 | 120 | 160 | 240t | 240s | 480 | |
| S_1A1 | S_1A1 | S_1A1 | S_1A1 | S_1A1 | S_1A1 | S_1A1 | 1 |
| S_1A2 | S_1A2 | S_1A2 | S_1A2 | | | F7_1A1 | 2 |
| F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | 3 |
| F1_1A2 | F1_1A2 | F1_1A2 | F1_1A2 | | | F8_1A1 | 4 |
| | F2_1A1 | F2_1A1 | F2_1A1 | F2_1A1 | F2_1A1 | F2_1A1 | 5 |
| | F2_1A2 | F2_1A2 | F2_1A2 | | | F9_1A1 | 6 |
| | | F3_1A1 | F3_1A1 | F3_1A1 | F3_1A1 | F3_1A1 | 7 |
| | | F3_1A2 | F3_1A2 | | | F10_1A1 | 8 |
| | | | F4_1A1 | F4_1A1 | F4_1A1 | F4_1A1 | 9 |
| | | | F4_1A2 | | | | 10 |
| S_1A4 | | | | F5_1A1 | F5_1A1 | F5_1A1 | 11 |
| F1_1A4 | | | | | | F11_1A1 | 12 |
| S_1A3 | | | | F6_1A1 | F6_1A1 | F6_1A1 | 13 |
| F1_1A3 | | | | | | F12_1A1 | 14 |

Aggregator 1-4

| Configuration | | | | | | | | | | Aggregator Input bus |
|---------------|--------|--------|--------|---------|---------|---------|--|--|--|----------------------------|
| 40 | 80 | 120 | 160 | 240f | 240s | 480 | | | | |
| F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | | | | 1 |
| F1_1A2 | F1_1A2 | F1_1A2 | F1_1A2 | | | F1_25A1 | | | | 2 |
| F1_1B1 | F1_1B1 | F1_1B1 | F1_1B1 | F1_1B1 | F1_1B1 | F1_1B1 | | | | 3 |
| F1_1B2 | F1_1B2 | F1_1B2 | F1_1B2 | | | F1_25B1 | | | | 4 |
| F1_1C1 | F1_1C1 | F1_1C1 | F1_1C1 | F1_1C1 | F1_1C1 | F1_1C1 | | | | 5 |
| F1_1C2 | F1_1C2 | F1_1C2 | F1_1C2 | | | F1_25C1 | | | | 6 |
| F1_1D1 | F1_1D1 | F1_1D1 | F1_1D1 | F1_1D1 | F1_1D1 | F1_1D1 | | | | 7 |
| F1_1D2 | F1_1D2 | F1_1D2 | F1_1D2 | | | F1_25D1 | | | | 8 |
| F1_1A4 | | F1_9A1 | F1_9A1 | F1_9A1 | F1_9A1 | F1_9A1 | | | | 9 |
| | | F1_9A2 | F1_9A2 | | | F1_33A1 | | | | 10 |
| F1_1B4 | | F1_9B1 | F1_9B1 | F1_9B1 | F1_9B1 | F1_9B1 | | | | 11 |
| | | F1_9B2 | F1_9B2 | | | F1_33B1 | | | | 12 |
| F1_1C4 | | F1_9C1 | F1_9C1 | F1_9C1 | F1_9C1 | F1_9C1 | | | | 13 |
| | | F1_9C2 | F1_9C2 | | | F1_33C1 | | | | 14 |
| F1_1D4 | | F1_9D1 | F1_9D1 | F1_9D1 | F1_9D1 | F1_9D1 | | | | 15 |
| | | F1_9D2 | F1_9D2 | | | F1_33D1 | | | | 16 |
| F1_1A3 | | | | F1_17A1 | F1_17A1 | F1_17A1 | | | | 17 |
| | | | | | | F1_41A1 | | | | 18 |
| F1_1B3 | | | | F1_17B1 | F1_17B1 | F1_17B1 | | | | 19 |
| | | | | | | F1_41B1 | | | | 20 |
| F1_1C3 | | | | F1_17C1 | F1_17C1 | F1_17C1 | | | | 21 |
| | | | | | | F1_41C1 | | | | 22 |
| F1_1D3 | | | | F1_17D1 | F1_17D1 | F1_17D1 | | | | 23 |
| | | | | | | F1_41D1 | | | | 24 |

Aggregator 5-8

| Configuration | | | | | | | Aggregator Input bus |
|---------------|--------|--------|---------|---------|---------|---------|----------------------------|
| 40 | 80 | 120 | 160 | 240f | 240s | 480 | |
| | F1_5A1 | F1_5A1 | F1_5A1 | F1_5A1 | F1_5A1 | F1_5A1 | 1 |
| | F1_5A2 | F1_5A2 | F1_5A2 | | | F1_29A1 | 2 |
| | F1_5B1 | F1_5B1 | F1_5B1 | F1_5B1 | F1_5B1 | F1_5B1 | 3 |
| | F1_5B2 | F1_5B2 | F1_5B2 | | | F1_29B1 | 4 |
| | F1_5C1 | F1_5C1 | F1_5C1 | F1_5C1 | F1_5C1 | F1_5C1 | 5 |
| | F1_5C2 | F1_5C2 | F1_5C2 | | | F1_29C1 | 6 |
| | F1_5D1 | F1_5D1 | F1_5D1 | F1_5D1 | F1_5D1 | F1_5D1 | 7 |
| | F1_5D2 | F1_5D2 | F1_5D2 | | | F1_29D1 | 8 |
| | | | F1_13A1 | F1_13A1 | F1_13A1 | F1_13A1 | 9 |
| | | | F1_13A2 | | | F1_37A1 | 10 |
| | | | F1_13B1 | F1_13B1 | F1_13B1 | F1_13B1 | 11 |
| | | | F1_13B2 | | | F1_37B1 | 12 |
| | | | F1_13C1 | F1_13C1 | F1_13C1 | F1_13C1 | 13 |
| | | | F1_13C2 | | | F1_37C1 | 14 |
| | | | F1_13D1 | F1_13D1 | F1_13D1 | F1_13D1 | 15 |
| | | | F1_13D2 | | | F1_37D1 | 16 |
| | | | | F1_21A1 | F1_21A1 | F1_21A1 | 17 |
| | | | | F1_21B1 | F1_21B1 | F1_45A1 | 18 |
| | | | | | | F1_21B1 | 19 |
| | | | | | | F1_45B1 | 20 |
| | | | | | | F1_21C1 | 21 |
| | | | | | | F1_45C1 | 22 |
| | | | | | | F1_21D1 | 23 |
| | | | | | | F1_45D1 | 24 |

121

Journal of Management Studies, 19(1), 67-80.

Separator 1-4

| Configuration | | | | | | | Separator output bus |
|---------------|--------|--------|--------|--------|---------|---------|----------------------------|
| 40 | 80 | 120 | 160 | 240f | 240s | 480 | |
| F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | 1 |
| F1_1A2 | F1_1A2 | F1_1A2 | F1_1A2 | F1_1A2 | | F1_25A1 | 2 |
| F1_1B1 | F1_1B1 | F1_1B1 | F1_1B1 | F1_1B1 | F1_1B1 | F1_1B1 | 3 |
| F1_1B2 | F1_1B2 | F1_1B2 | F1_1B2 | F1_1B2 | | F1_25B1 | 4 |
| F1_1C1 | F1_1C1 | F1_1C1 | F1_1C1 | F1_1C1 | F1_1C1 | F1_1C1 | 5 |
| F1_1C2 | F1_1C2 | F1_1C2 | F1_1C2 | F1_1C2 | | F1_25C1 | 6 |
| F1_1D1 | F1_1D1 | F1_1D1 | F1_1D1 | F1_1D1 | F1_1D1 | F1_1D1 | 7 |
| F1_1D2 | F1_1D2 | F1_1D2 | F1_1D2 | F1_1D2 | | F1_25D1 | 8 |
| F1_1A4 | | F1_9A1 | F1_9A1 | F1_9A1 | F1_9A1 | F1_9A1 | 9 |
| F1_1A5 | | F1_9A2 | F1_9A2 | F1_9A2 | | F1_33A1 | 10 |
| F1_1B4 | | F1_9B1 | F1_9B1 | F1_9B1 | F1_9B1 | F1_9B1 | 11 |
| F1_1B5 | | F1_9B2 | F1_9B2 | F1_9B2 | | F1_33B1 | 12 |
| F1_1C4 | | F1_9C1 | F1_9C1 | F1_9C1 | F1_9C1 | F1_9C1 | 13 |
| F1_1C5 | | F1_9C2 | F1_9C2 | F1_9C2 | | F1_33C1 | 14 |
| F1_1D4 | | F1_9D1 | F1_9D1 | F1_9D1 | F1_9D1 | F1_9D1 | 15 |
| F1_1D5 | | F1_9D2 | F1_9D2 | | | F1_33D1 | 16 |
| F1_1A3 | F1_1A3 | F1_1A3 | | | F1_17A1 | F1_17A1 | 17 |
| F1_1B3 | F1_1B3 | F1_1B3 | | | F1_17A2 | F1_41A1 | 18 |
| | | F1_9B3 | | | F1_17B1 | F1_17B1 | 19 |
| F1_1C3 | F1_1C3 | F1_1C3 | | | F1_17B2 | F1_41B1 | 20 |
| | | F1_9C3 | | | F1_17C1 | F1_17C1 | 21 |
| F1_1D3 | F1_1D3 | F1_1D3 | | | F1_17C2 | F1_41C1 | 22 |
| | | F1_9D3 | | | F1_17D1 | F1_17D1 | 23 |
| | | | | | F1_17D2 | F1_41D1 | 24 |

Separator 5-8

| Configuration | | | | | | | | Separator output bus |
|---------------|--------|--------|---------|---------|---------|---------|--|----------------------------|
| 40 | 80 | 120 | 160 | 240t | 240s | 480 | | |
| | F1_5A1 | F1_5A1 | F1_5A1 | F1_5A1 | F1_5A1 | F1_5A1 | | 1 |
| | F1_5A2 | F1_5A2 | F1_5A2 | F1_5A2 | | F1_29A1 | | 2 |
| | F1_5B1 | F1_5B1 | F1_5B1 | F1_5B1 | F1_5B1 | F1_5B1 | | 3 |
| | F1_5B2 | F1_5B2 | F1_5B2 | F1_5B2 | | F1_29B1 | | 4 |
| | F1_5C1 | F1_5C1 | F1_5C1 | F1_5C1 | F1_5C1 | F1_5C1 | | 5 |
| | F1_5C2 | F1_5C2 | F1_5C2 | F1_5C2 | | F1_29C1 | | 6 |
| | F1_5D1 | F1_5D1 | F1_5D1 | F1_5D1 | F1_5D1 | F1_5D1 | | 7 |
| | F1_5D2 | F1_5D2 | F1_5D2 | F1_5D2 | | F1_29D1 | | 8 |
| | | | F1_13A1 | F1_13A1 | F1_13A1 | F1_13A1 | | 9 |
| | | | F1_13A2 | F1_13A2 | | F1_37A1 | | 10 |
| | | | F1_13B1 | F1_13B1 | F1_13B1 | F1_13B1 | | 11 |
| | | | F1_13B2 | F1_13B2 | | F1_37B1 | | 12 |
| | | | F1_13C1 | F1_13C1 | F1_13C1 | F1_13C1 | | 13 |
| | | | F1_13C2 | F1_13C2 | | F1_37C1 | | 14 |
| | | | F1_13D1 | F1_13D1 | F1_13D1 | F1_13D1 | | 15 |
| | | | F1_13D2 | F1_13D2 | | F1_37D1 | | 16 |
| | F1_5A3 | F1_5A3 | | F1_21A1 | F1_21A1 | F1_21A1 | | 17 |
| | | | | F1_21A2 | | F1_45A1 | | 18 |
| | F1_5B3 | F1_5B3 | | F1_21B1 | F1_21B1 | F1_21B1 | | 19 |
| | | | | F1_21B2 | | F1_45B1 | | 20 |
| | F1_5C3 | F1_5C3 | | F1_21C1 | F1_21C1 | F1_21C1 | | 21 |
| | | | | F1_21C2 | | F1_45C1 | | 22 |
| | F1_5D3 | F1_5D3 | | F1_21D1 | F1_21D1 | F1_21D1 | | 23 |
| | | | | F1_21D2 | | F1_45D1 | | 24 |

Separator 9

| Configuration | | | | | | | Separator output bus |
|---------------|----------|----------|----------|----------|----------|----------|----------------------------|
| 40 | 80 | 120 | 160 | 240t | 240s | 480 | |
| F1_CP_A1 | F1_CP_A1 | F1_CP_A1 | F1_CP_A1 | F1_CP_A1 | F1_CP_A1 | F1_CP_A1 | 1 |
| F1_CP_A2 | F1_CP_A2 | F1_CP_A2 | F1_CP_A2 | F1_CP_A2 | | | 2 |
| F1_CP_B1 | F1_CP_B1 | F1_CP_B1 | F1_CP_B1 | F1_CP_B1 | F1_CP_B1 | F1_CP_B1 | 3 |
| F1_CP_B2 | F1_CP_B2 | F1_CP_B2 | F1_CP_B2 | F1_CP_B2 | | | 4 |
| F1_CP_C1 | F1_CP_C1 | F1_CP_C1 | F1_CP_C1 | F1_CP_C1 | F1_CP_C1 | F1_CP_C1 | 5 |
| F1_CP_C2 | F1_CP_C2 | F1_CP_C2 | F1_CP_C2 | F1_CP_C2 | | | 6 |
| F1_CP_D1 | F1_CP_D1 | F1_CP_D1 | F1_CP_D1 | F1_CP_D1 | F1_CP_D1 | F1_CP_D1 | 7 |
| F1_CP_D2 | F1_CP_D2 | F1_CP_D2 | F1_CP_D2 | F1_CP_D2 | | | 8 |
| F1_CP_A4 | | | | | | | 9 |
| F1_CP_A5 | | | | | | | 10 |
| F1_CP_B4 | | | | | | | 11 |
| F1_CP_B5 | | | | | | | 12 |
| F1_CP_C4 | | | | | | | 13 |
| F1_CP_C5 | | | | | | | 14 |
| F1_CP_D4 | | | | | | | 15 |
| F1_CP_D5 | | | | | | | 16 |
| F1_CP_A3 | F1_CP_A3 | F1_CP_A3 | | | | | 17 |
| F1_CP_B3 | F1_CP_B3 | F1_CP_B3 | | | | | 18 |
| F1_CP_C3 | F1_CP_C3 | F1_CP_C3 | | | | | 19 |
| F1_CP_D3 | F1_CP_D3 | F1_CP_D3 | | | | | 20 |
| | | | | | | | 21 |
| | | | | | | | 22 |
| | | | | | | | 23 |
| | | | | | | | 24 |

125

Unstriper

| Configuration | | | | | | | Unstriper Input bus |
|---------------|--------|--------|--------|--------|--------|---------|---------------------------|
| 40 | 80 | 120 | 160 | 240f | 240s | 480 | |
| S_1A1 | S_1A1 | S_1A1 | S_1A1 | S_1A1 | S_1A1 | S_1A1 | 1 |
| S_1A2 | S_1A2 | S_1A2 | S_1A2 | S_1A2 | | F7_1A1 | 2 |
| F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | F1_1A1 | 3 |
| F1_1A2 | F1_1A2 | F1_1A2 | F1_1A2 | F1_1A2 | | F8_1A1 | 4 |
| | F2_1A1 | F2_1A1 | F2_1A1 | F2_1A1 | F2_1A1 | F2_1A1 | 5 |
| | F2_1A2 | F2_1A2 | F2_1A2 | F2_1A2 | | F9_1A1 | 6 |
| S_1A5 | | F3_1A1 | F3_1A1 | F3_1A1 | F3_1A1 | F3_1A1 | 7 |
| F1_1A5 | | F3_1A2 | F3_1A2 | F3_1A2 | | F10_1A1 | 8 |
| S_1A4 | | | F4_1A1 | F4_1A1 | F4_1A1 | F4_1A1 | 9 |
| F1_1A4 | | | F4_1A2 | F4_1A2 | | | 10 |
| S_1A3 | S_1A3 | S_1A3 | | F5_1A1 | F5_1A1 | F5_1A1 | 11 |
| F1_1A3 | F1_1A3 | F1_1A3 | | F5_1A2 | | F11_1A1 | 12 |
| | F2_1A3 | F2_1A3 | | F6_1A1 | F6_1A1 | F6_1A1 | 13 |
| | | F3_1A3 | | F6_1A2 | | F12_1A1 | 14 |

09040113 063000

gigabit signals

Switch size
number of fabrics
base data bus

| | | | | | | |
|----|----|-----|-----|------|------|-----|
| 40 | 80 | 120 | 160 | 240f | 240s | 480 |
| 1 | 2 | 3 | 4 | 6 | 6 | 12 |
| 12 | 6 | 4 | 3 | 2 | 2 | 1 |

-- gigabits
-- does not include spare
-- number of 250MHz data bits from striper to aggregator per OC-48 (4x per port card)

Calculated
stripers to fabric bus

| | | | | | | | |
|----------------------|------|------|------|----|------|------|-----|
| data in gigabits | 26 | 14 | 10 | 8 | 6 | 6 | 4 |
| transmitters | 3.25 | 1.75 | 1.25 | 1 | 0.75 | 0.75 | 0.5 |
| receivers | 4 | 2 | 2 | 1 | 1 | 1 | 1 |
| data out gigabits | 52 | 28 | 20 | 16 | 12 | 12 | 8 |
| receivers | 6.5 | 3.5 | 2.5 | 2 | 1.5 | 1.5 | 1 |
| | 7 | 4 | 3 | 2 | 2 | 2 | 1 |

@125MHz to Tx
/8 bits per Tx
round up to unit Tx
@125MHz from Rx (w/2x speed-up)
/8 bits per Rx
round up to unit Rx

| | | | | | | | | |
|------------------------|--------------|------|------|------|------|------|------|------|
| stripers to fabric bus | transmitters | 4 | 2 | 2 | 2 | 1 | 1 | 1 |
| receivers | 5 | 3 | 3 | 2 | 2 | 1 | 1 | 1 |
| speedup | 1.50 | 1.67 | 2.00 | 2.00 | 2.00 | 1.00 | 2.00 | 2.00 |

-- update these numbers to affect input speedup
-- update these numbers to affect output speedup
speed-up assumes always use 4 RW bits in egress direction

Configs

40/80 80/120 120/160 160/240f 240f/240s 240s/480

required simultaneous modes

| | | | | | | | | |
|----------------|-----------|----|----|----|----|----|----|----|
| Port Card | active Tx | 10 | 8 | 10 | 12 | 7 | 13 | 13 |
| stripers | active Rx | 13 | 12 | 14 | 14 | 14 | 13 | 14 |
| unique streams | | | | | | | | 56 |

max active Tx number of serializers per striper
max active Rx number of deserializers per stripers
number of quads transceivers per port card

| | | | | | | | | |
|---------------|--------------|----|---|---|---|---|---|-----|
| total streams | Tx increment | 10 | 2 | 2 | 2 | 0 | 6 | 22 |
| stripers | Rx increment | 13 | 3 | 2 | 4 | 0 | 6 | 28 |
| unstripers | | | | | | | | 200 |

number of gigabit backplane connections per stripers
number of gigabit backplane connections per unstripers
number of gigabit backplane connections per port card
number of pins per port card

| | | | | | | | | |
|--------|-----------|----|----|----|----|----|----|-----|
| Fabric | active Rx | 16 | 16 | 16 | 20 | 12 | 24 | 24 |
| Agg1-4 | active Rx | 8 | 8 | 16 | 20 | 12 | 24 | 24 |
| Agg5-8 | active Tx | 20 | 24 | 24 | 24 | 24 | 24 | 24 |
| Sep1-4 | active Tx | 12 | 12 | 20 | 24 | 24 | 24 | 24 |
| Sep5-8 | | | | | | | | 192 |

max active Rx number of deserializers per odd aggregator
max active Rx number of deserializers per even aggregator
max active Tx number of serializers per odd separator
max active Tx number of serializers per even separator
number of transceivers per fabric
number of quad transceivers per fabric

| | | | | | | | | |
|---------------|--------------|----|----|---|---|---|----|-----|
| total streams | Rx increment | 16 | 8 | 0 | 4 | 0 | 12 | 40 |
| Agg1-4 | Rx increment | 8 | 0 | 8 | 4 | 0 | 12 | 32 |
| Agg5-8 | Tx increment | 20 | 12 | 0 | 8 | 0 | 12 | 52 |
| Sep1-4 | Tx increment | 12 | 0 | 8 | 8 | 0 | 12 | 40 |
| Sep5-8 | | | | | | | | 656 |

number of gigabit backplane connections per odd aggregator
number of gigabit backplane connections per even aggregator
number of gigabit backplane connections per odd separator
number of gigabit backplane connections per even separator
number of gigabit backplane connections per fabric
number of pins per fabric